SIGIR 2002 Workshops
Thursday August 15, 2002, Pinni Building, University of Tampere

2. INFORMATION RETRIEVAL AND OCR: FROM CONVERTING CONTENT TO GRASPING MEANING (Auditorium 2107)

Organizers

Jamie Callan, CMU
David Grossman, IIT
Alex Hauptmann, CMU
Paul Kantor, Rutgers

Description of the topic

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 on 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 world wide access to millions of pages of legacy print documents, representing the shared human record as it existed until just a few years ago.

This workshop will 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 will 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.

The workshop will welcome 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.

Program

AM Invited/Solicited Overview papers

10.00 **1. Introduction. Links to other groups, especially Document Recognition and Retrieval Conference X**
Paul Kantor (Rutgers University)

Abstract: Applications ranging from conversion of legacy document, to scanning moving images, to detecting criminal activity in FAX messages bring new importance to the issue of organizing and retrieval on poorly imaged documents. Cognate conferences provide other perspectives, and some recent advances presented elsewhere will be summarized, to stimulate cross-fertilization.

10.40 **2. Retrieving OCR Text: A Survey of Current Approaches**

Steven M. Beitzel, Eric C. Jensen & David A. Grossman (Information Retrieval Laboratory. Department of Computer Science Illinois Institute of Technology)

Abstract: The importance of effectively retrieving OCR text has grown significantly in recent years. We provide a brief overview of work done to improve the effectiveness of retrieval of OCR text.

11.20 **3. Discussion: Define themes and challenges for the breakouts.**

12.00 Lunch: Breakout groups discussing the themes and challenges.

PM

1. Contributed Papers

13.30 **1.1 A Voting System for Automatic OCR Correction**

S.T. Klein & M. Kopel (Bar Ilan University)

Abstract: The present work suggests a new approach to improve the quality of OCR output. It is based on the simple, yet surprising, fact that different OCR devices tend to make different mistakes. We found many errors which were sometimes frequent for one device but rare for the other, and vice versa. This suggests applying two or more OCR algorithms and matching their outputs. The presented post-processing method is applied automatically, without consulting the user. By that we succeeded to achieve a more accurate and more efficient method for correcting OCR generated errors.

14.00 **1.2 A Clustering-Based Algorithm for Automatic Document Separation**

Kevyn Collins-Thompson (Language Technologies Institute School of Computer Science, Carnegie Mellon University)

Radoslav Nickolov (Microsoft Corporation, USA)

Abstract: For text, audio, video, and still images, a number of projects have addressed the problem of estimating inter-object similarity and the related problem of finding transition, or 'segmentation' points in a stream of objects of the same media type. There has been relatively little work in this area for document images, which are typically text-intensive and contain a mixture of layout, text-based, and image features. Beyond simple partitioning, the problem of clustering related page images is also important, especially for information retrieval problems such as document image searching and browsing. Motivated by this, we describe a model for estimating inter-page similarity in ordered collections of document images, based on a combination of text and layout features. The features are used as input to a discriminative classifier, whose output is used in a constrained clustering criterion. We do a task-based evaluation of our method by applying it the problem of automatic document separation during batch scanning.

14.30 **1.3. A Content-based Probabilistic Correction Model for OCR Document Retrieval**

Rong Jin, Alex G. Hauptmann & ChengXiang Zhai (School of Computer Science, Carnegie Mellon University)

Abstract: The difficulty with information retrieval for OCR documents lies in the fact that OCR documents comprise of a significant amount of erroneous words and unfortunately most information retrieval techniques rely heavily on word matching between documents and queries. In this paper, we propose a general content-based correction model that can work on top of an existing OCR correction tool to "boost" retrieval performance. The basic idea of this correction model is to exploit the whole content of a document to supplement any other useful information provided by an existing OCR correction tool for word correction. Instead of making an explicit correction decision for each erroneous word as typically done in a traditional approach, we consider the uncertainties in such correction decisions and compute an estimate of the original "uncorrupted" document language model accordingly. The document language model can then be used for retrieval with a language modelling retrieval approach. Evaluation using a TREC standard testing collection indicates that our method significantly improves the performance when compared with simple word correction approaches such as using only the top ranked correction.

15.00 **1.3 Examining the Effectiveness of IR Techniques for Document Image Retrieval** Gareth J. F. Jones & Adenike M. Lam-Adesina (Department of Computer Science, University of Exeter)

Abstract: Effective Document Image Retrieval (DIR) requires the use of appropriate Information Retrieval (IR) methods. Our research indicates that some IR techniques found to be effective for electronic text retrieval do not transfer to DIR in a simple predictable manner, and that modifications must be made to these methods to enable them work effectively for DIR.

15.30 **2. Summary discussion/reports from lunch discussions**