Journal of Universal Computer Science, vol. 17, no. 1 (2011), 1-2
submitted: 17/12/10, accepted: 28/12/10, appeared: 1/1/11 © J.UCS

Meeting New Challenges in Document Engineering

J.UCS Special Issue

Rafael Dueire Lins
(Universidade Federal de Pernambuco, Recife, Brazil
rdl.ufpe@gmail.com)

A document is any object that "carries" information. This wide scope definition allows one to see from pre-historic painting in a cave wall to a 3-D film as documents. The most usual kind of document is a "paper" document. Electronic ones, such as pdf files, are becoming of widespread use today. Document engineering is the area of knowledge concerned with principles, tools and processes that improve our ability to create, manage, store, compact, access, and maintain documents. The fields of document recognition and retrieval have grown rapidly in recent years. Such development has been fueled by the emergence of new application areas such as the World Wide Web (WWW), digital libraries, and video- and camera-based OCR.

The main areas of concern in Document Engineering are:

- Algorithms and systems for machine-printed and handwritten character and word recognition, especially for degraded documents (e.g., faxes);
- Character and word segmentation techniques;
- Identification and analysis of tables or equations;
- Page segmentation, including hierarchical decomposition of documents into text regions, halftones, colored/textured background, etc;
- Logical/linguistic structure analysis and recognition of documents
- Raster-to-vector conversion of line-art, maps, and technical drawings;
- Document image filtering, enhancement and compression techniques;
- Document degradation models;
- Video and camera based OCR;
- Applications of document recognition to the WWW and digital libraries;
- Techniques to support spoken language access to document text;
- Multilingual character recognition;
- Impact of recognition accuracy on retrieval effectiveness;
- Recovery and use of logical structure for retrieval;
- Relevance feedback techniques for document retrieval;
- Cross-language and multi-lingual retrieval;
- Categorization and summarization of text documents and image documents;
- Keyword spotting in document images;
- Approximate string matching algorithms for OCRs;
- Non-textual retrieval methods;
- Image and multimedia search;
- Interfaces for document retrieval

Contents of this Issue

Originally, 13 submissions from 12 different countries, spread over Africa, the Americas, Asia and Europe were received in response to the call-for-papers. A board of specialists of all areas of document engineering carefully reviewed all submissions and selected ten papers covering different areas of the field. The authors revised all papers according to referees´ recommendations and re-submitted their contribution, which were forwarded to the original set of reviewers. The final versions of the approved papers appear here. Now, the ten papers are shortly overviewed.

This volume opens with a classical problem in document engineering in the paper "Document Retrieval Using SIFT Image Features", by Dan Smith and Richard Harvey from the University of East Anglia, UK.

Ergina Kavallieratou and Fotis Daskas from Greece addressed the important problem of "Text Line Detection and Segmentation: Uneven Skew Angles and Hill-and-Dale Writing". Along the same research line, Darko Brodić and Zoran Milivojević from Serbia contributed with the paper "A New Approach to Water Flow Algorithm for Text Line Segmentation".

OCR, Optical Character Recognition, is a processing intensive procedure, overall when one has to find where the text lies. To try to make such a task more efficient, a group of researchers from Brazil presented the interesting paper "An OCR Free Method for Word Spotting in Printed Documents: the Evaluation of Different Feature Sets". "The Use of Latent Semantic Indexing to Mitigate OCR Effects of Related Document Images", by researchers from Spain and Brazil, follows a different path and tries to enhance OCR performance.

The study of non Latin-based writing systems is a hot topic in the area of document engineering. From Bangalore, India, this volume received the paper by R. Rampalli and A. G. Ramakrishnan "Fusion of Complementary Online and Offline Strategies for Recognition of Handwritten Kannada Characters" and also the article "Choice of Classifiers in Hierarchical Recognition of Online Handwritten Kannada and Tamil Aksharas" by N. Venkatesh and A. G. Ramakrishnan. Also from Bangalore, comes the paper "Color Image RestorationUsing Neural Network Model", by S. Chickerur and A. Kumar.

Daniela da Cruz and Pedro Rangel Henriques, from Universidade do Minho, Portugal contributed to this volume with the paper "Visualizing and Analyzing the Quality of XML Documents".

This volume closes with a report by me entitled "Nabuco – Two Decades of Document Processing in Latin America", which details a pioneer initiative in building a library of historical documents.

Acknowledgements

The editor and the authors of this volume are grateful for the patience and enthusiasm of Prof. Dr. Hermann Maurer and Dana Kaiser that made it viable and to the large number of reviewers that anonymous and carefully refereed each paper submitted.

Rafael Dueire Lins
Guest editor
Recife (Brazil), December 2010