Journal of Universal Computer Science, vol. 14, no. 2 (2008), 174-177
submitted: 4/1/08, accepted: 17/1/08, appeared: 28/1/08

Advances in Document Engineering

J.UCS Special Issue

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

Document Engineering is a discipline within computer science that investigates systems for documents in any form and in all media. Document engineering is 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. This 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 use of OCR is spreading from high-volume, niche domains to more general tasks, including the processing of noisy "real-world" documents, photocopies, and faxes.

These are the main areas of concern in Document Engineering:

- 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 structure analysis and recognition, linguistic representation of document structure;
- 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 (audio browsing of doc. databases);
- 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;
- Benchmarking and evaluation issues

Contents of this Issue

The start up for this volume was The Document Engineering track of ACM-SAC 2007 that was held in Seoul (Korea) from March 11 to 15, 2007. Originally, 26 submissions from 12 different countries, spread over Africa, the Americas, Asia and Europe were received. A board of 58 specialists reviewed all submissions and selected eight papers and two posters, covering different areas of the field. All papers and posters were invited to be resubmitted to this special issue. Authors had the freedom to revise, detail and update their final submissions. Nine of them were approved in the final versions that appear herein.

Two papers report on extracting content information of documents. Sylvain Lamprier and collaborators from LERIA – Université Angers (France) present an algorithm for semantic linear text segmentation on general corpuses. A new specialist tool for medieval document XML markup is introduced by Georg Vogeler (Germany), Benjamin Buckard (Germany), and Stefan Gruner (South Africa).

Character recognition is the focus of the paper presented by Cinthia Freitas and her colleagues from Pontifical Catholic University of Paraná and Universidade Tecnológica Federal do Paraná (Brazil) entitled "Metaclasses and Zoning Mechanism Applied to Handwriting Recognition". In the same research theme there is also the paper "A Progressive Learning Method for Symbol Recognition" by Sabine Barrat and Salvatore Tabbone from LORIA - Université Nancy 2 (France). The interesting problem of signature authenticity is addressed in the paper "Combining Classifiers in the ROC-space for Off-line Signature Verification" by Luiz Oliveira, Edson Justino and Flavio Bortolozzi from Brazil together with Robert Sabourin from Canada.

Image enhancement and segmentation is the target of the four other selected papers. The paper entitled "Table-form Extraction with Artifact Removal" by Luiz Neves, João Carvalho, Jacques Facon and Flávio Bortolozzi, from PUC-Paraná (Brazil) presents a novel methodology for extracting the structure of handwritten filled table-forms. The three other papers address the removal of back-to-front interference of documents written on both sides of translucent paper. The article "Detailing a Quantitative Method for Assessing Algorithms to Remove Back-to-Front Interference in Documents" is offered by me and João Marcelo da Silva from Brazil together with F. Mário Martins from UMINHO (Braga – Portugal). Wafa Boussellaa and Adnan Amin from University of Sfax (Tunisia) together with Abdezarrak Zahour from University of Le Havre (France) focused on removing back-to-front interference in Arabic historical documents in their article "A Methodology for the Separation of Foreground/Background in Arabic Historical Manuscripts using Hybrid Methods". João Marcelo da Silva and I (UFPE – Brazil) together with F. Mário Martins from UMINHO (Braga – Portugal) and Rosita Wachenchauzer from UBA – Argentina are

presenting “A New and Efficient Algorithm to Binarize Document Images Removing Back-to-Front Interference”.

Reviewers

The experts of all areas of document engineering from all over the world that composed the program committee of the ACM-SAC 2007 track on Document Engineering and also revised the papers for this special issue are presented below:

Adel M. Alimi *(University of Sfax, Tunisia)*
Angelo Marcelli *(University of Salerno, Italy)*
Apóstolos Antonacopoulos *(Univ. of Salford,UK)*
Alejandro C. Frery *(Univ. Federal de Alagoas, Brazil)*
Andreas Dengel *(Kaiserslautern Univ., Germany)*
Antony Wiley *(Hewlett Packard Labs., Bristol, UK)*
Aurélio Campilho *(Univ. do Porto, Portugal)*
Daniel P. Lopresti *(Lehigh University, USA)*
David S. Doermann *(University of Maryland, USA)*
Dov Dori *(Technion, Israel Inst. of Technology, Israel)*
Ethan Munson *(Univ. of Wisconsin – Milwaukee, USA)*
Flávio Bortolozzi *(P. Univ. Católica do Paraná, Brazil)*
Graham Leedham *(Nanyang Tech. University, Singapore)*
Henry S. Baird *(Lehigh University, USA)*
Hirobumi Nishida *(Ricoh Sw Research Center, Japan)*
Horst Bunke *(University of Bern, Switzerland)*
Jacques Facon *(P. Univ. Católica do Paraná, Brazil)*
Jin H. Kim *(Computer Science Department, Korea)*
Jian Liang (*Media Management Tech, USA*)
João Marques de Carvalho (*UF Campina Grande, Brazil*)
Jonathan J. Hull. *(Ricoh Calif. Research Center, USA)*
Josep Llados *(Univ. Autonoma de Barcelona, Spain)*
Kazem Taghva *(University of Nevada, USA)*
Lawrence O'Gorman *(Avaya Labs, USA)*
Luis Corte-Real *(Univ. do Porto, Portugal)*
Louisa Lam *(Hong Kong Inst. of Education, Hong Kong)*
Majid Mirmehdi *(University of Bristol, England)*
Marco Gori *(Università di Siena, Italy)*
Maria Feldgen *(Univ. Buenos Aires, Argentina)*
Michael Perrone *(IBM T.J. Watson Research Center, USA)*
Mohamed Kamel *(Univ. of Waterloo, Canada)*
Nasser Sherkat *(The Nottingham Trent University, England)*
Nelson Mascarenhas *(Universidade de São Paulo, Brazil)*
Pedro Rangel Henriques *(U. do Minho em Braga, Portugal)*
Pertti Vakkari. *(University of Tampere, Finland)*
Rafael Dueire Lins *(U.F. Pernambuco, Brazil)*
Ricardo de Queiroz *(Univ. de Brasília, Brazil)*

Acknowledgements

The editor and the authors of this volume are grateful for the enthusiasm of Prof. Dr. Hermann Maurer and Dana Kaiser that made it possible.

Recife (Brazil), January 2008