Lecture Notes in Computer Science 3163

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

David Hutchison
Lancaster University, UK

Takeo Kanade
Carnegie Mellon University, Pittsburgh, PA, USA

Josef Kittler
University of Surrey, Guildford, UK

Jon M. Kleinberg
Cornell University, Ithaca, NY, USA

Friedemann Mattern
ETH Zurich, Switzerland

John C. Mitchell
Stanford University, CA, USA

Moni Naor
Weizmann Institute of Science, Rehovot, Israel

Oscar Nierstrasz
University of Bern, Switzerland

C. Pandu Rangan
Indian Institute of Technology, Madras, India

Bernhard Steffen
University of Dortmund, Germany

Madhu Sudan
Massachusetts Institute of Technology, MA, USA

Demetri Terzopoulos
New York University, NY, USA

Doug Tygar
University of California, Berkeley, CA, USA

Moshe Y. Vardi
Rice University, Houston, TX, USA

Gerhard Weikum
Max-Planck Institute of Computer Science, Saarbruecken, Germany

Simone Marinai   Andreas Dengel (Eds.)

Document Analysis Systems VI

6th International Workshop, DAS 2004
Florence, Italy, September 8 - 10, 2004
Proceedings

Springer

Volume Editors

Simone Marinai
Università di Firenze, Dipartimento di Sistemi e Informatica
Via S. Marta, 3 - 50139 Firenze, Italy
E-mail: marinai@dsi.unifi.it

Andreas Dengel
German Research Center for Artificial Intelligence (DFKI)
P.O.Box 2080, 67608 Kaiserslautern, Germany
E-mail: Andreas.Dengel@dfki.de

Library of Congress Control Number: 2004111168

CR Subject Classification (1998): I.5, H.3, I.4, I.7, J.1, J.2

ISSN 0302-9743
ISBN 3-540-23060-2 Springer Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

Springer is a part of Springer Science+Business Media

springeronline.com

© Springer-Verlag Berlin Heidelberg 2004
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Olgun Computergrafik
Printed on acid-free paper SPIN: 11321026 06/3142 5 4 3 2 1 0

Preface

This volume contains papers selected for presentation at the 6th IAPR Workshop on Document Analysis Systems (DAS 2004) held during September 8–10, 2004 at the University of Florence, Italy. Several papers represent the state of the art in a broad range of "traditional" topics such as layout analysis, applications to graphics recognition, and handwritten documents. Other contributions address the description of complete working systems, which is one of the strengths of this workshop. Some papers extend the application domains to other media, like the processing of Internet documents.

The peculiarity of this 6th workshop was the large number of papers related to digital libraries and to the processing of historical documents, a taste which frequently requires the analysis of color documents. A total of 17 papers are associated with these topics, whereas two years ago (in DAS 2002) only a couple of papers dealt with these problems.

In our view there are three main reasons for this new wave in the DAS community. From the scientific point of view, several research fields reached a thorough knowledge of techniques and problems that can be effectively solved, and this expertise can now be applied to new domains. Another incentive has been provided by several research projects funded by the EC and the NSF on topics related to digital libraries. Last but not least, the organization of focused events, like the recent DIAL workshop chaired by Henry Baird and Venu Govindraju in Palo Alto (CA), had a strong impact on the definition of new research directions. However, it is indeed a lucky coincidence that this new trend in DAS research emerged in this edition organized in a town such as Florence, which keeps such an exceptional artistic and cultural heritage.

We received a total of 79 submissions from 19 countries, and we selected 31 oral presentations and 22 posters highlighted with short oral introductions. As a supplement to this proceedings, notes from the workshop discussions and other material related to presented papers will be posted on the DAS 2004 website: http://www.dsi.unifi.it/DAS04. Each paper was reviewed by three reviewers whom we would like to warmly thank here. We should mention the valuable support and hints provided by members of the Program Committee and past DAS chairs. We also wish to acknowledge the generosity of our sponsors: the International Association for Pattern Recognition, the University of Florence, the DFKI, ABBYY, Hitachi, and Siemens.

Special thanks are due to Alessio Ceroni, Cristina Dolfi, and Emanuele Marino for their invaluable contributions to the local organization.

June 2004

Simone Marinai
Andreas Dengel

Organization

Workshop Co-chairs

Simone Marinai	University of Florence, Italy
Andreas Dengel	DFKI, Germany

Program Committee

Apostolos Antonacopoulos	University of Liverpool, UK
Henry Baird	Lehigh University, USA
Francesca Cesarini	University of Florence, Italy
David Doermann	University of Maryland, USA
Andrew Downton	University of Essex, UK
Hiromichi Fujisawa	Hitachi Central Research Laboratory, Japan
Jianying Hu	IBM T.J. Watson Research Center, USA
Rolf Ingold	University of Fribourg, Switzerland
Ramanujan Kashi	Avaya Labs Research, USA
Koichi Kise	Osaka Prefecture University, Japan
Dan Lopresti	Lehigh University, USA
Donato Malerba	University of Bari, Italy
Udo Miletzki	Siemens Dematic, Germany
Yasuaki Nakano	Kyushu University, Japan
Lambert Schomaker	Rijksuniversiteit Groningen, The Netherlands
Giovanni Soda	University of Florence, Italy
Larry Spitz	Document Recognition Technologies, New Zealand
Karl Tombre	LORIA-INPL, France
Luc Vincent	LizardTech, USA
Marcel Worring	University of Amsterdam, The Netherlands

Additional Referees

Annalisa Appice
Margherita Berardi
Alain Biem
Thomas Breuel
Michelangelo Ceci
Philippe Dosch
Stefan Jaeger
Dimosthenis Karatzas
Michele Lapi
Larry O'Gorman
Huanfeng Ma
Gérald Masini
Eugene Ratzlaff
Maurizio Rigamonti
T.R. Roth-Berghofer
Jane Snowdon
Salvatore Tabbone
Yefeng Zheng
Gary Zi

Table of Contents

Digital Libraries

Document Analysis Systems for Digital Libraries:
Challenges and Opportunities 1
Henry S. Baird, Venugopal Govindaraju, and Daniel P. Lopresti

The Trinity College Dublin 1872 Online Catalogue 17
John G. Byrne

DL Architecture for Indic Scripts 28
Suryaprakash Kompalli, Srirangaraj Setlur, and Venugopal Govindaraju

A Semantic-Based System for Querying Personal Digital Libraries 39
Luigi Cinque, Alessio Malizia, and Roberto Navigli

Toward Personalized Digital Library for Providing "Information JIT" 47
*Hisashi Ikeda, Naohiro Furukawa, Katsumi Marukawa,
and Hiromichi Fujisawa*

Historical Documents

Tilting at Windmills:
Adventures in Attempting to Reconstruct *Don Quixote* 51
A. Lawrence Spitz

A Segmentation-Free Recognition Technique
to Assist Old Greek Handwritten Manuscript OCR 63
*Basilios Gatos, Kostas Ntzios, Ioannis Pratikakis, Sergios Petridis,
T. Konidaris, and Stavros J. Perantonis*

Automatic Metadata Retrieval from Ancient Manuscripts 75
Frank Le Bourgeois and Hala Kaileh

A Complete Approach to the Conversion
of Typewritten Historical Documents for Digital Archives 90
Apostolos Antonacopoulos and Dimosthenis Karatzas

An Adaptive Binarization Technique
for Low Quality Historical Documents 102
Basilios Gatos, Ioannis Pratikakis, and Stavros J. Perantonis

Segmentation of Handwritten Characters
for Digitalizing Korean Historical Documents 114
Min Soo Kim, Kyu Tae Cho, Hee Kue Kwag, and Jin Hyung Kim

Self-organizing Maps and Ancient Documents 125
Eddie Smigiel, Abdel Belaid, and Hatem Hamza

Enriching Historical Manuscripts: The Bovary Project 135
Stéphane Nicolas, Thierry Paquet, and Laurent Heutte

Layout Analysis

Word Grouping in Document Images Based on Voronoi Tessellation 147
Yue Lu, Zhe Wang, and Chew Lim Tan

Multi-component Document Image Coding Using Regions-of-Interest 158
Xiao Wei Yin, Andy C. Downton, Martin Fleury, and J. He

Physical Layout Analysis of Complex Structured Arabic Documents
Using Artificial Neural Nets 170
Karim Hadjar and Rolf Ingold

An Integrated Approach for Automatic Semantic Structure Extraction
in Document Images 179
Margherita Berardi, Michele Lapi, and Donato Malerba

Multi-view HAC for Semi-supervised Document Image Classification 191
Fabien Carmagnac, Pierre Héroux, and Éric Trupin

Configurable Text Stamp Identification Tool
with Application of Fuzzy Logic 201
J. He and Andy C. Downton

Layout and Content Extraction for PDF Documents 213
Hui Chao and Jian Fan

Automatic Extraction of Filled-In Items from Bank-Check Images 225
Katsuhiko Ueda, Hirotoshi Maegawa, and Kenichi Matsuo

Color Documents

Bleed-Through Removal from Degraded Documents
Using a Color Decorrelation Method 229
Anna Tonazzini, Emanuele Salerno, Matteo Mochi, and Luigi Bedini

Colour Map Classification for Archive Documents 241
J. He and Andy C. Downton

Serialized k-Means for Adaptative Color Image Segmentation –
Application to Document Images and Others 252
Yann Leydier, Frank Le Bourgeois, and Hubert Emptoz

Adaptive Region Growing Color Segmentation for Text Using Irregular Pyramid 264
Poh Kok Loo and Chew Lim Tan

Preprocessing and Segmentation of Bad Quality Machine Typed Documents 276
Mariusz Szwoch and Wioleta Szwoch

Handwritten Documents

Ensembles of Classifiers for Handwritten Word Recognition Specialized on Individual Handwriting Style 286
Simon Günter and Horst Bunke

Information Retrieval System for Handwritten Documents 298
Sargur Srihari, Anantharaman Ganesh, Catalin Tomai, Yong-Chul Shin, and Chen Huang

Word–Wise Script Identification from Indian Documents 310
Suranjit Sinha, Umapada Pal, and B.B. Chaudhuri

Recognizing Freeform Digital Ink Annotations 322
Michael Shilman and Zile Wei

Post-processing of Handwritten Pitman's Shorthand Using Unigram and Heuristic Approaches 332
Swe Myo Htwe, Colin Higgins, Graham Leedham, and Ma Yang

Multiscale Handwriting Characterization for Writers' Classification 337
Véronique Eglin, Stéphane Bres, and Carlos Rivero

Graphics Recognition

A Hybrid Approach to Detect Graphical Symbols in Documents 342
Salvatore Tabbone, Laurent Wendling, and Daniel Zuwala

Performance Evaluation of Symbol Recognition 354
Ernest Valveny and Philippe Dosch

The Search for Genericity in Graphics Recognition Applications: Design Issues of the Qgar Software System 366
Jan Rendek, Gérald Masini, Philippe Dosch, and Karl Tombre

Attributed Graph Matching Based Engineering Drawings Retrieval 378
Rujie Liu, Takayuki Baba, and Daiki Masumoto

A Platform to Extract Knowledge from Graphic Documents. Application to an Architectural Sketch Understanding Scenario 389
Gemma Sánchez, Ernest Valveny, Josep Lladós, Joan Mas, and Narcís Lozano

Internet Documents

A Graph-Based Framework for Web Document Mining 401
Adam Schenker, Horst Bunke, Mark Last, and Abraham Kandel

XML Documents Within a Legal Domain:
Standards and Tools for the Italian Legislative Environment 413
Carlo Biagioli, Enrico Francesconi, Pierluigi Spinosa, and Mirco Taddei

Rule-Based Structural Analysis of Web Pages 425
Fabio Vitali, Angelo Di Iorio, and Elisa Ventura Campori

Extracting Table Information from the Web 438
Yeon-Seok Kim and Kyong-Ho Lee

A Neural Network Classifier for Junk E-Mail 442
Ian Stuart, Sung-Hyuk Cha, and Charles Tappert

Document Analysis Systems

Results of a Study on Invoice-Reading Systems in Germany 451
Bertin Klein, Stevan Agne, and Andreas Dengel

A Document Analysis System Based on Text Line Matching
of Multiple OCR Outputs 463
Yasuaki Nakano, Toshihiro Hananoi, Hidetoshi Miyao,
Minoru Maruyama, and Ken-ichi Maruyama

DocMining: A Document Analysis System Builder 472
Sébastien Adam, Maurizio Rigamonti, Eric Clavier, Éric Trupin,
Jean-Marc Ogier, Karl Tombre, and Joël Gardes

Automatic Fax Routing 484
Paul Viola, James Rinker, and Martin Law

Applications

Contextual *Swarm*-Based Multi-layered Lattices:
A New Architecture for Contextual Pattern Recognition 496
David G. Elliman and Sherin M. Youssef

Natural Language Processing of Patents and Technical Documentation . . . 508
Gaetano Cascini, Alessandro Fantechi, and Emilio Spinicci

Document Image Retrieval in a Question Answering System
for Document Images 521
Koichi Kise, Shota Fukushima, and Keinosuke Matsumoto

A Robust Braille Recognition System 533
Apostolos Antonacopoulos and David Bridson

Document Image Watermarking Based on Weight-Invariant Partition Using Support Vector Machine 546
Shiyan Hu

Video Degradation Model and Its Application to Character Recognition in e-Learning Videos 555
Jun Sun, Yutaka Katsuyama, and Satoshi Naoi

Unity Is Strength: Coupling Media for Thematic Segmentation 559
Dalila Mekhaldi, Denis Lalanne, and Rolf Ingold

Author Index 563