

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

Horst Bunke A. Lawrence Spitz (Eds.)

Document Analysis Systems VII

7th International Workshop, DAS 2006
Nelson, New Zealand, February 13-15, 2006
Proceedings



Springer

Volume Editors

Horst Bunke
University of Bern
Department of Computer Science
Neubrückstr. 10, 3012 Bern, Switzerland
E-mail: bunke@iam.unibe.ch

A. Lawrence Spitz
DocRec Ltd
34 Strathaven Place, Atawhai, Nelson 7001, New Zealand
E-mail: spitz@docrec.com

Library of Congress Control Number: 2005939178

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

LNCS Sublibrary: SL 6 – Image Processing, Computer Vision, Pattern Recognition, and Graphics

ISSN	0302-9743
ISBN-10	3-540-32140-3 Springer Berlin Heidelberg New York
ISBN-13	978-3-540-32140-8 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 2006
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11669487 06/3142 5 4 3 2 1 0

Preface

DAS 2006 is the Seventh International Association for Pattern Recognition Workshop on Document Analysis Systems and was held in Nelson, New Zealand. DAS 2006 built on the tradition of past workshops held in Kaiserslautern, Germany (1994), Malvern, PA (1996), Nagano, Japan (1998), Rio de Janeiro, Brazil (2000), Princeton, NJ (2002), and Florence, Italy (2004). The goal of this meeting was to bring together those who have designed systems, or systems components, to solve real-world problems in document analysis.

Document analysis systems is inherently an interdisciplinary field encompassing such diverse disciplines as image processing, pattern recognition, document structure and natural language processing. DAS 2006 attempted to bring these disciplines together and to provide interactions between systems developers, suppliers and end users.

We received 78 papers from 19 countries. Each submission was reviewed by three reviewers. In addition to the Program Committee members, 42 other reviewers helped in this process. From those submissions and their reviews, we went through the difficult and sometimes painful process of ranking papers for acceptance or rejection. In the end we accepted 33 papers for oral presentation and 22 for presentation at poster sessions.

We, the Co-chairmen of DAS 2006, wish to express our gratitude to all of our colleagues who have reviewed the papers submitted for this conference.

We are proud to have brought two distinguished speakers to Nelson for keynote addresses: Ian Witten of the University of Waikato, the father of the New Zealand Digital Library, and James Fruchterman, a pioneer in modern commercial optical character recognition and currently CEO of Benetech.

We owe a special debt of gratitude to Marcus Liwicki of the University of Bern for his tireless work at maintaining the website, managing the flow of papers and reviews into the ConfMan system and assembling the proceedings for publication by Springer. He was ably assisted by Andreas Schlapbach.

We are fortunate that Siemens, Hitachi and Humanware provided DAS with financial support, and we thank them for doing so. Additionally, following the DAS tradition, the organizers of DAS 2004 have passed on the surplus from running that workshop for our use.

But ultimately it is the collection of authors who submitted papers to DAS to whom we owe the greatest gratitude. It is on them and their high-quality submissions that the success of DAS 2006 relies.

February 2006

Horst Bunke and Larry Spitz
Program Chairs
DAS 2006

Organization

DAS 2006 was organized by DocRec Ltd.

Executive Committee

Conference Chairs: Larry Spitz (DocRec Ltd, New Zealand)
Horst Bunke (University of Bern, Switzerland)

Program Committee

Apostolos Antonacopoulos (UK)
Henry Baird (USA)
Thomas Breuel (Germany)
Horst Bunke (Switzerland)
Andreas Dengel (Germany)
David Doermann (USA)
Andrew Downton (UK)
Michael Fairhurst (UK)
Hiromichi Fujisawa (Japan)
Venugopal Govindaraju (USA)
Tin Kam Ho (USA)
Jianying Hu (USA)
Rolf Ingold (Switzerland)
Rangachar Kasturi (USA)
Koichi Kise (Japan)
Seong-Whan Lee (Korea)
Daniel Lopresti (USA)
Raghavan Manmatha (USA)
Simone Marinai (Italy)
Udo Miletzki (Germany)
Yasuaki Nakano (Japan)
Larry Spitz (New Zealand)
Karl Tombre (France)

Referees

Stefan Agne
Andrew Bagdanov
Ardhendu Behera
Koustav Bhattacharya
Alain Biem

Jean-Luc Blöchle
Matthew Boonstra
Jakob Brendel
Joshua Candamo
Farzin Deravi

VIII Organization

Faisal Farooq
Gunnar Grimnes
Richard Guest
Sanaul Hoque
Gareth Howells
Jonathan Hull
Masakazu Iwamura
Stefan Jaeger
Thomas Kieninger
Malte Kiesel
Bertin Klein
Dar-Shyang Lee
Hansheng Lei
Jian Liang
Rainer Lindwurm
Vasant Manohar
Dalila Mekhaldi

David Mihalcik
Tristan Miller
Pranab Mohanty
Sunita Nayak
Shinichiro Omachi
Christoph Pesch
Maurizio Rigamonti
Thomas Roth-Berghofer
Sven Schwarz
Karthik Sridharan
Seiichi Uchida
Himanshu Vajaria
Ludger van Elst
Shankar Vembu
Alan Yang

Sponsoring Institutions

Siemens AG, Munich, Germany
HumanWare Group, Christchurch, New Zealand
Hitachi Central Research Laboratory, Tokyo, Japan

Scientific Sponsors

DocRec Ltd, Atawhai, Nelson, New Zealand
University of Bern, Switzerland
International Association for Pattern Recognition

Table of Contents

Session 1: Digital Libraries

Retrieval from Document Image Collections <i>A. Balasubramanian, Million Meshesha, C.V. Jawahar</i>	1
A Semi-automatic Adaptive OCR for Digital Libraries <i>Sachin Rawat, K.S. Sesh Kumar, Million Meshesha, Indraneel Deb Sikdar, A. Balasubramanian, C.V. Jawahar</i>	13

Session 2: Image Processing

Contribution to the Discrimination of the Medieval Manuscript Texts: Application in the Palaeography <i>Ikram Moalla, Frank LeBourgeois, Hubert Emptoz, Adel M. Alimi</i>	25
Restoring Ink Bleed-Through Degraded Document Images Using a Recursive Unsupervised Classification Technique <i>Drira Fadoua, Frank Le Bourgeois, Hubert Emptoz</i>	38
Networked Document Imaging with Normalization and Optimization <i>Hirobumi Nishida</i>	50
Gray-Scale Thinning Algorithm Using Local Min/Max Operations <i>Kyoung Min Kim, Buham Lee, Nam Sup Choi, Gwan Hee Kang, Joong Jo Park, Ching Y. Suen</i>	62

Session 3: Handwriting 1

Automated Scoring of Handwritten Essays Based on Latent Semantic Analysis <i>Sargur Srihari, Jim Collins, Rohini Srihari, Pavithra Babu, Harish Srinivasan</i>	71
Aligning Transcripts to Automatically Segmented Handwritten Manuscripts <i>Jamie Rothfeder, R. Manmatha, Toni M. Rath</i>	84
Virtual Example Synthesis Based on PCA for Off-Line Handwritten Character Recognition <i>Hidetoshi Miyao, Minoru Maruyama</i>	96

Extraction of Handwritten Text from Carbon Copy Medical Form Images <i>Robert Milewski, Venu Govindaraju</i>	106
---	-----

Session 4: Document Structure and Format

Document Logical Structure Analysis Based on Perceptive Cycles <i>Yves Rangoni, Abdel Belaïd</i>	117
A System for Converting PDF Documents into Structured XML Format <i>Hervé Déjean, Jean-Luc Meunier</i>	129
XCDF: A Canonical and Structured Document Format <i>Jean-Luc Bloechle, Maurizio Rigamonti, Karim Hadjar, Denis Lalanne, Rolf Ingold</i>	141
Structural Analysis of Mathematical Formulae with Verification Based on Formula Description Grammar <i>Seiichi Toyota, Seiichi Uchida, Masakazu Suzuki</i>	153

Session 5: Tables

Notes on Contemporary Table Recognition <i>David W. Embley, Daniel Lopresti, George Nagy</i>	164
Handwritten Artefact Identification Method for Table Interpretation with Little Use of Previous Knowledge <i>Luiz Antônio Pereira Neves, João Marques de Carvalho, Jacques Facon, Flávio Bortolozzi, Sérgio Aparecido Ignácio</i>	176

Session 6: Handwriting 2

Writer Identification for Smart Meeting Room Systems <i>Marcus Liwicki, Andreas Schlapbach, Horst Bunke, Samy Bengio, Johnny Mariéthoz, Jonas Richiardi</i>	186
Extraction and Analysis of Document Examiner Features from Vector Skeletons of Grapheme ‘th’ <i>Vladimir Pervouchine, Graham Leedham</i>	196
Segmentation of On-Line Handwritten Japanese Text Using SVM for Improving Text Recognition <i>Bilan Zhu, Junko Tokuno, Masaki Nakagawa</i>	208

Application of Bi-gram Driven Chinese Handwritten Character Segmentation for an Address Reading System <i>Yan Jiang, Xiaoqing Ding, Qiang Fu, Zheng Ren</i>	220
--	-----

Session 7: Language and Script Identification

Language Identification in Degraded and Distorted Document Images <i>Shijian Lu, Chew Lim Tan, Weihua Huang</i>	232
Bangla/English Script Identification Based on Analysis of Connected Component Profiles <i>Lijun Zhou, Yue Lu, Chew Lim Tan</i>	243
Script Identification from Indian Documents <i>Gopal Datt Joshi, Saurabh Garg, Jayanthi Sivaswamy</i>	255
Finding the Best-Fit Bounding-Boxes <i>Bo Yuan, Leong Keong Kwoh, Chew Lim Tan</i>	268

Session 9: Systems and Performance Evaluation

Towards Versatile Document Analysis Systems <i>Henry S. Baird, Matthew R. Casey</i>	280
Exploratory Analysis System for Semi-structured Engineering Logs <i>Michael Flaster, Bruce Hillyer, Tin Kam Ho</i>	291
Ground Truth for Layout Analysis Performance Evaluation <i>A. Antonacopoulos, D. Karatzas, D. Bridson</i>	302
On Benchmarking of Invoice Analysis Systems <i>Bertin Klein, Stefan Agne, Andreas Dengel</i>	312
Semi-automatic Ground Truth Generation for Chart Image Recognition <i>Li Yang, Weihua Huang, Chew Lim Tan</i>	324

Session 10: Retrieval and Segmentation

Efficient Word Retrieval by Means of SOM Clustering and PCA <i>Simone Marinai, Stefano Faini, Emanuele Marino, Giovanni Soda</i> ...	336
The Effects of OCR Error on the Extraction of Private Information <i>Kazem Taghva, Russell Beckley, Jeffrey Coombs</i>	348

Combining Multiple Classifiers for Faster Optical Character Recognition <i>Kumar Chellapilla, Michael Shilman, Patrice Simard</i>	358
--	-----

Performance Comparison of Six Algorithms for Page Segmentation <i>Faisal Shafait, Daniel Keysers, Thomas M. Breuel</i>	368
---	-----

Posters

HVS Inspired System for Script Identification in Indian Multi-script Documents <i>Peeta Basa Pati, A.G. Ramakrishnan</i>	380
---	-----

A Shared Fragments Analysis System for Large Collections of Web Pages <i>Junchang Ma, Zhimin Gu</i>	390
--	-----

Offline Handwritten Arabic Character Segmentation with Probabilistic Model <i>Pingping Xiu, Liangrui Peng, Xiaoqing Ding, Hua Wang</i>	402
---	-----

Automatic Keyword Extraction from Historical Document Images <i>Kengo Terasawa, Takeshi Nagasaki, Toshio Kawashima</i>	413
---	-----

Digitizing a Million Books: Challenges for Document Analysis <i>K. Pramod Sankar, Vamshi Ambati, Lakshmi Pratha, C.V. Jawahar</i>	425
--	-----

Toward File Consolidation by Document Categorization <i>Abdel Belaïd, André Alusse</i>	437
---	-----

Finding Hidden Semantics of Text Tables <i>Saleh A. Alrashed</i>	449
---	-----

Reconstruction of Orthogonal Polygonal Lines <i>Alexander Gribov, Eugene Bodansky</i>	462
--	-----

A Multiclass Classification Framework for Document Categorization <i>Qi Qiang, Qinming He</i>	474
--	-----

The Restoration of Camera Documents Through Image Segmentation <i>Shijian Lu, Chew Lim Tan</i>	484
---	-----

Cut Digits Classification with k-NN Multi-specialist <i>Fernando Boto, Andoni Cortés, Clemente Rodríguez</i>	496
---	-----

The Impact of OCR Accuracy and Feature Transformation on Automatic Text Classification <i>Mayo Murata, Lazaro S.P. Busagala, Wataru Ohyama, Tetsushi Wakabayashi, Fumitaka Kimura</i>	506
A Method for Symbol Spotting in Graphical Documents <i>Daniel Zuwala, Salvatore Tabbone</i>	518
Groove Extraction of Phonographic Records <i>Sylvain Stotzer, Ottar Johnsen, Frédéric Bapst, Rolf Ingold</i>	529
Use of Affine Invariants in Locally Likely Arrangement Hashing for Camera-Based Document Image Retrieval <i>Tomohiro Nakai, Koichi Kise, Masakazu Iwamura</i>	541
Robust Chinese Character Recognition by Selection of Binary-Based and Grayscale-Based Classifier <i>Yoshinobu Hotta, Jun Sun, Yutaka Katsuyama, Satoshi Naoi</i>	553
Segmentation-Driven Recognition Applied to Numerical Field Extraction from Handwritten Incoming Mail Documents <i>Clément Chatelain, Laurent Heutte, Thierry Paquet</i>	564
Performance Evaluation of Text Detection and Tracking in Video <i>Vasant Manohar, Padmanabhan Soundararajan, Matthew Boonstra, Harish Raju, Dmitry Goldgof, Rangachar Kasturi, John Garofolo</i>	576
Document Analysis System for Automating Workflows <i>Steven J. Simske, Jordi Arnabat</i>	588
Automatic Assembling of Cadastral Maps Based on Generalized Hough Transformation <i>Fei Liu, Wataru Ohyama, Tetsushi Wakabayashi, Fumitaka Kimura</i>	593
A Few Steps Towards On-the-Fly Symbol Recognition with Relevance Feedback <i>Jan Rendek, Bart Lamiroy, Karl Tombre</i>	604
The Fuzzy-Spatial Descriptor for the Online Graphic Recognition: Overlapping Matrix Algorithm <i>Noorazrin Zakaria, Jean-Marc Ogier, Josep Lladós</i>	616
Author Index	629