Horst Bunke  A. Lawrence Spitz (Eds.)

Document Analysis Systems VII

7th International Workshop, DAS 2006
Proceedings

Springer
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
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

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
   A. Balasubramanian, Million Meshesha, C.V. Jawahar ............... 1

A Semi-automatic Adaptive OCR for Digital Libraries
   Sachin Rawat, K.S. Sesh Kumar, Million Meshesha,
   Indraneel Deb Sikdar, A. Balasubramanian, C.V. Jawahar ........... 13

Session 2: Image Processing

Contribution to the Discrimination of the Medieval Manuscript Texts:
Application in the Palaeography
   Ikram Moalla, Frank LeBourgeois, Hubert Emptoz, Adel M. Alimi .... 25

Restoring Ink Bleed-Through Degraded Document Images Using a
Recursive Unsupervised Classification Technique
   Drina Fadoua, Frank Le Bourgeois, Hubert Emptoz ................. 38

Networked Document Imaging with Normalization and Optimization
   Hirobumi Nishida .................................................... 50

Gray-Scale Thinning Algorithm Using Local Min/Max Operations
   Kyoung Min Kim, Buhm Lee, Nam Sup Choi, Gwan Hee Kang,
   Joong Jo Park, Ching Y. Suen .................................. 62

Session 3: Handwriting 1

Automated Scoring of Handwritten Essays Based on Latent Semantic
Analysis
   Sargur Srihari, Jim Collins, Rohini Srihari, Pavithra Babu,
   Harish Srinivasan .................................................. 71

Aligning Transcripts to Automatically Segmented Handwritten
Manuscripts
   Jamie Rothfeder, R. Manmatha, Toni M. Rath ...................... 84

Virtual Example Synthesis Based on PCA for Off-Line Handwritten
Character Recognition
   Hidetoshi Miyao, Minoru Maruyama ................................ 96
Extraction of Handwritten Text from Carbon Copy Medical Form Images
   Robert Milewski, Venu Govindaraju ............................................ 106

Session 4: Document Structure and Format

Document Logical Structure Analysis Based on Perceptive Cycles
   Yves Rangoni, Abdel Belaïd .................................................. 117

A System for Converting PDF Documents into Structured XML Format
   Hervé Déjean, Jean-Luc Meunier ............................................ 129

XCDF: A Canonical and Structured Document Format
   Jean-Luc Bloechle, Maurizio Rigamonti, Karim Hadjar,
    Denis Lalanne, Rolf Ingold .................................................. 141

Structural Analysis of Mathematical Formulae with Verification Based on Formula Description Grammar
   Seiichi Toyota, Seiichi Uchida, Masakazu Suzuki .................... 153

Session 5: Tables

Notes on Contemporary Table Recognition
   David W. Embley, Daniel Lopresti, George Nagy ....................... 164

Handwritten Artefact Identification Method for Table Interpretation with Little Use of Previous Knowledge
   Luiz Antônio Pereira Neves, João Marques de Carvalho,
    Jacques Facon, Flávio Bortolozzi, Sérgio Aparecido Ignácio ........ 176

Session 6: Handwriting 2

Writer Identification for Smart Meeting Room Systems
   Marcus Liwicki, Andreas Schlapbach, Horst Bunke, Samy Bengio,
    Johnny Mariéthoz, Jonas Richiardi ...................................... 186

Extraction and Analysis of Document Examiner Features from Vector Skeletons of Grapheme ‘th’
   Vladimir Pervouchine, Graham Leedham .................................. 196

Segmentation of On-Line Handwritten Japanese Text Using SVM for Improving Text Recognition
   Bilan Zhu, Junko Tokuno, Masaki Nakagawa ................................. 208
Session 7: Language and Script Identification

Language Identification in Degraded and Distorted Document Images
Shijian Lu, Chew Lim Tan, Weihua Huang ................................. 232

Bangla/English Script Identification Based on Analysis of Connected Component Profiles
Lijun Zhou, Yue Lu, Chew Lim Tan ........................................ 243

Script Identification from Indian Documents
Gopal Datt Joshi, Saurabh Garg, Jayanthi Sivaswamy .................. 255

Finding the Best-Fit Bounding-Boxes
Bo Yuan, Leong Keong Kwoh, Chew Lim Tan ......................... 268

Session 9: Systems and Performance Evaluation

Towards Versatile Document Analysis Systems
Henry S. Baird, Matthew R. Casey ........................................ 280

Exploratory Analysis System for Semi-structured Engineering Logs
Michael Flaster, Bruce Hillyer, Tin Kam Ho ........................... 291

Ground Truth for Layout Analysis Performance Evaluation
A. Antonacopoulos, D. Karatzas, D. Bridson ............................ 302

On Benchmarking of Invoice Analysis Systems
Bertin Klein, Stefan Agne, Andreas Dengel ............................. 312

Semi-automatic Ground Truth Generation for Chart Image Recognition
Li Yang, Weihua Huang, Chew Lim Tan ................................. 324

Session 10: Retrieval and Segmentation

Efficient Word Retrieval by Means of SOM Clustering and PCA
Simone Marinai, Stefano Faini, Emanuele Marino, Giovanni Soda 336

The Effects of OCR Error on the Extraction of Private Information
Kazem Taghva, Russell Beckley, Jeffrey Coombs ........................ 348
## Table of Contents

### Combining Multiple Classifiers for Faster Optical Character Recognition

*Kumar Chellapilla, Michael Shilman, Patrice Simard* .......................... 358

### Performance Comparison of Six Algorithms for Page Segmentation

*Faisal Shafait, Daniel Keysers, Thomas M. Breuel* ............................. 368

### Posters

**HVS Inspired System for Script Identification in Indian Multi-script Documents**

*Peeta Basa Pati, A.G. Ramakrishnan* ........................................... 380

**A Shared Fragments Analysis System for Large Collections of Web Pages**

*Junchang Ma, Zhimin Gu* ......................................................... 390

**Offline Handwritten Arabic Character Segmentation with Probabilistic Model**

*Pingping Xiu, Liangrui Peng, Xiaoqing Ding, Hua Wang* .................... 402

**Automatic Keyword Extraction from Historical Document Images**

*Kengo Terasawa, Takeshi Nagasaki, Toshio Kawashima* ........................ 413

**Digitizing a Million Books: Challenges for Document Analysis**

*K. Pramod Sankar, Vamshi Ambati, Lakshmi Pratha, C.V. Jawahar* ......... 425

**Toward File Consolidation by Document Categorization**

*Abdel Belaïd, André Alusse* ..................................................... 437

**Finding Hidden Semantics of Text Tables**

*Saleh A. Alrashed* ................................................................. 449

**Reconstruction of Orthogonal Polygonal Lines**

*Alexander Gribov, Eugene Bodansky* ......................................... 462

**A Multiclass Classification Framework for Document Categorization**

*Qi Qiang, Qinming He* ........................................................... 474

**The Restoration of Camera Documents Through Image Segmentation**

*Shijian Lu, Chew Lim Tan* ...................................................... 484

**Cut Digits Classification with k-NN Multi-specialist**

*Fernando Boto, Andoni Cortés, Clemente Rodríguez* ....................... 496
The Impact of OCR Accuracy and Feature Transformation on Automatic Text Classification
Mayo Murata, Lazaro S.P. Busagala, Wataru Ohyama, Tetsushi Wakabayashi, Fumitaka Kimura ........................................ 506

A Method for Symbol Spotting in Graphical Documents
Daniel Zuwala, Salvatore Tabbone ........................................ 518

Groove Extraction of Phonographic Records
Sylvain Stotzer, Ottar Johnsen, Frédéric Bapst, Rolf Ingold ............ 529

Use of Affine Invariants in Locally Likely Arrangement Hashing for Camera-Based Document Image Retrieval
Tomohiro Nakai, Koichi Kise, Masakazu Iwamura ........................ 541

Robust Chinese Character Recognition by Selection of Binary-Based and Grayscale-Based Classifier
Yoshinobu Hotta, Jun Sun, Yutaka Katsuyama, Satoshi Naoi ........... 553

Segmentation-Driven Recognition Applied to Numerical Field Extraction from Handwritten Incoming Mail Documents
Clément Chatelain, Laurent Heutte, Thierry Paquet ........................ 564

Performance Evaluation of Text Detection and Tracking in Video
Vasant Manohar, Padmanabhan Soundararajan, Matthew Boonstra, Harish Raju, Dmitry Goldgof, Rangachar Kasturi, John Garofolo .... 576

Document Analysis System for Automating Workflows
Steven J. Simske, Jordi Arnabat ............................................. 588

Automatic Assembling of Cadastral Maps Based on Generalized Hough Transformation
Fei Liu, Wataru Ohyama, Tetsushi Wakabayashi, Fumitaka Kimura ........................................ 593

A Few Steps Towards On-the-Fly Symbol Recognition with Relevance Feedback
Jan Rendek, Bart Lamiroy, Karl Tombre .................................. 604

The Fuzzy-Spatial Descriptor for the Online Graphic Recognition: Overlapping Matrix Algorithm
Noorazrin Zakaria, Jean-Marc Ogier, Josep Llados .......................... 616

Author Index ................................................................. 629