

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

Drira 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

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
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 <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