

Proceedings

Second International Conference on Document Image Analysis for Libraries

April 27 – 28, 2006
Lyon, France



Los Alamitos, California

Washington • Tokyo

All rights reserved.

Copyright and Reprint Permissions: Abstracting is permitted with credit to the source. Libraries may photocopy beyond the limits of US copyright law, for private use of patrons, those articles in this volume that carry a code at the bottom of the first page, provided that the per-copy fee indicated in the code is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923.

Other copying, reprint, or republication requests should be addressed to: IEEE Copyrights Manager, IEEE Service Center, 445 Hoes Lane, P.O. Box 133, Piscataway, NJ 08855-1331.

The papers in this book comprise the proceedings of the meeting mentioned on the cover and title page. They reflect the authors' opinions and, in the interests of timely dissemination, are published as presented and without change. Their inclusion in this publication does not necessarily constitute endorsement by the editors, the IEEE Computer Society, or the Institute of Electrical and Electronics Engineers, Inc.

IEEE Computer Society Order Number P2531

ISBN 0-7695-2531-8

Library of Congress Number 2006923027

Additional copies may be ordered from:

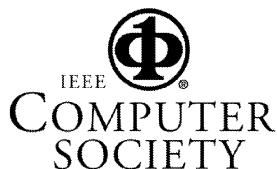
IEEE Computer Society
Customer Service Center
10662 Los Vaqueros Circle
P.O. Box 3014
Los Alamitos, CA 90720-1314
Tel: + 1 800 272 6657
Fax: + 1 714 821 4641
<http://computer.org/cspress>
csbooks@computer.org

IEEE Service Center
445 Hoes Lane
P.O. Box 1331
Piscataway, NJ 08855-1331
Tel: + 1 732 981 0060
Fax: + 1 732 981 9667
[http://shop.ieee.org/store/
customer-service@ieee.org](http://shop.ieee.org/store/customer-service@ieee.org)

IEEE Computer Society
Asia/Pacific Office
Watanabe Bldg., 1-4-2
Minami-Aoyama
Minato-ku, Tokyo 107-0062
JAPAN
Tel: + 81 3 3408 3118
Fax: + 81 3 3408 3553
tokyo.ofc@computer.org

Individual paper REPRINTS may be ordered at: <reprints@computer.org>

Editorial production by Bob Werner
Cover art production by Joe Daigle/Studio Productions
Printed in the United States of America by The Printing House



IEEE Computer Society
Conference Publishing Services
<http://www.computer.org/proceedings/>

Table of Contents: DIAL 2006

Second International Conference on Document Image Analysis for Libraries

Preface	viii
Conference Organization	ix
Introduction Paper	
Interactive Document Processing and Digital Libraries <i>George Nagy and Daniel Lopresti</i>	2
Chapter 1. Handwritten Documents Segmentation and Recognition	
Separating Lines of Text in Free-Form Handwritten Historical Documents <i>Douglas J. Kennard and William A. Barrett</i>	12
Multi-Queue Merging Scheme and its Application in Arabic Script Segmentation <i>Pingping Xiu, Liangrui Peng, and Xiaoqing Ding</i>	24
Exploring the Use of Conditional Random Field Models and HMMs for Historical Handwritten Document Recognition <i>Shaolei Feng, R. Manmatha, and Andrew McCallum</i>	30
Text Line Extraction in Handwritten Document with Kalman Filter Applied on Low Resolution Image <i>Aur�lie Lemaitre and Jean Camillerapp</i>	38
Complex Handwritten Page Segmentation Using Contextual Models <i>St�phane Nicolas, Thierry Paquet, and Laurent Heutte</i>	46
Chapter 2. OCR and Layout Analysis for Multilingual Printed Documents	
Dedicated Texture Based Tools for Characterisation of Old Books <i>Nicholas Journet, V�ronique Eglin, Jean-Yves Ramel, and R�my Mullot</i>	60
Detection and Segmentation of Table of Contents and Index Pages from Document Images <i>Sekhar Mandal, Shyama P. Chowdhury, Amit K. Das, and Bhabatosh Chanda</i>	70
Improving OCR Text Categorization Accuracy with Electronic Abstracts <i>Linlin Li and Chew Lim Tan</i>	82
Character Templates Learning for Textual Images Recognition as an Example of Learning in Structural Recognition <i>Bogdan Savchynskyy and Olexander Kamotskyy</i>	88
Design and Comparison of Segmentation Driven and Recognition Driven Devanagari OCR <i>Suryaprakash Kompalli, Srirangaraj Setlur, and Venu Govindaraju</i>	96
Combining a Hybrid Approach for Features Selection and Hidden Markov Models in Multifont Arabic Characters Recognition <i>Nadia Ben Amor and Najoua Essoukri Ben Amara</i>	103

OCR Voting Methods for Recognizing Low Contrast Printed Documents _____	108
<i>István Marosi, László Tóth</i>	

Chapter 3. DIA Applications for Digital Libraries

The Case of the Digitized Works at a National Digital Library _____	116
<i>José Borbinha, João Gil, Gilberto Pedrosa, and João Penas</i>	
Multilingual Document Recognition Research and its Application in China _____	126
<i>Liangrui Peng, Changsong Liu, Xiaoqing Ding, and Hua Wang</i>	
Digital Bleaching and Content Extraction for the Digital Archive of Rare Books _____	133
<i>Asanobu Kitamoto, Takeo Yamamoto, Makiko Onishi, Tomohiro Ikezaki, Ryo Kamida, Dominique Deuff, Eka Meyer, Sonoko Sato, Takako Muramatsu, and Kinji Ono</i>	
AGORA: The Interactive Document Image Analysis Tool of the BVH Project _____	145
<i>Jean Yves Ramel, Sébastien Busson, and Marie-Luce Demonet</i>	
Persée: Addressing the Needs of the Digitalisation and Online Accessibility of Back Collections through Robust and Integrated Tools _____	156
<i>Bruno Morandière and Viviane Boulétreau</i>	
User Centered Image Management System for Digital Libraries _____	164
<i>Zoltán Iszlai and Előd Egyed-Zsigmond</i>	
What Can We Learn from the Processing of 165,000 Forms from the 19th Century? _____	172
<i>Bertrand Couasnon</i>	

Chapter 4. Documents Images Indexing and Retrieval

Use of Figures in Literature Mining for Biomedical Digital Libraries _____	180
<i>Nawei Chen, Hagit Shatkay, and Dorothea Blostein</i>	
Document Image Retrieval Using Signatures as Queries _____	198
<i>Sargur N. Srihari, Shravya Shetty, Siyuan Chen, Harish Srinivasan, Chen Huang, Gady Agam, and Ophir Frieder</i>	
Automatic Content-Based Indexing of Digital Documents through Intelligent Processing Techniques _____	204
<i>Floriana Esposito, Stefano Ferilli, Teresa M.A. Basile, and Nicola Di Mauro</i>	
On Defining Signatures for the Retrieval and the Classification of Graphical Drop Caps _____	220
<i>Rudolf Pareti, Surapong Uttama, Jean-Pierre Salmon, Jean-Marc Ogier, Salvatore Tabbone, Laurent Wendling, Sebastien Adam, and Nicole Vincent</i>	
Distance Measures for Layout-Based Document Image Retrieval _____	232
<i>Joost van Beusekom, Daniel Keysers, Faisal Shafait, and Thomas M. Breuel</i>	
Tree Clustering for Layout-Based Document Image Retrieval _____	243
<i>Simone Marinai, Emanuele Marino, and Giovanni Soda</i>	

Chapter 5. Other Trends in DIA

Ink Recognition Based on Statistical Classification Methods _____	254
<i>Vassiliki Kokla, Alexandra Psarrou, and Vassilis Konstantinou</i>	
Computer Assistance for Digital Libraries: Contributions to Middle-Ages and Authors' Manuscripts Exploitation and Enrichment _____	265
<i>Veronique Eglin, Frank LeBourgeois, Stephane Bres, Hubert Emptoz, Yann Leydier, Ikram Moalla, and Fadoua Drira</i>	
Querying for Silhouettes by Qualitative Feature Schemes _____	281
<i>Björn Gottfried</i>	
A Novel Technique for the Watermarking of Symbolically Compressed Documents _____	291
<i>Sarbani Palit and Utpal Garain</i>	
Refinement of Digitized Documents through Recognition of Mathematical Formulae _____	297
<i>Toshihiro Kanahori and Masakazu Suzuki</i>	
Image Analysis for Palaeography Inspection _____	303
<i>Ikram Moalla, Frank Le Bourgeois, Hubert Emptoz, and Adel M. Alimi</i>	

Chapter 6. DIA Evaluation and Document Image Quality

Quality Assurance in High Volume Document Digitization: A Survey _____	312
<i>Xiaofan Lin</i>	
Performance Evaluation of a Mathematical Formula Recognition System with a Large Scale of Printed Formula Images _____	320
<i>Kazuki Ashida, Masayuki Okamoto, and Hiroki Imai</i>	
Experimental Performance Study of a User Intensive and Large-scale Digital Library Framework _____	332
<i>Xiangwen Liao, Binxing Fang, Weihua Luo, and Bin Wang</i>	
Improving the Quality of Degraded Document Images _____	340
<i>Ergina Kavallieratou, Efstathios Stamatatos</i>	
Towards Restoring Historic Documents Degraded Over Time _____	350
<i>Fadoua Drira</i>	
Image Interpolation Using Mathematical Morphology _____	358
<i>Alessandro Ledda, Hiệp Q. Luong, Wilfried Philips, Valérie De Witte, and Etienne E. Kerre</i>	
An Image Inpainting Approach Based on the Poisson Equation _____	368
<i>Xiaowei Shao, Zhengkai Liu, and Houqiang Li</i>	
DIAL 2004 Working Group Report on Acquisition Quality Control _____	373
<i>Elisa H. Barney Smith, Henry Baird, William Barrett, Frank Le Bourgeois, Xiaofan Lin, George Nagy, and Steve Simske</i>	
Author Index _____	377