

Document Recognition and Retrieval XVIII

Gady Agam
Christian Viard-Gaudin
Editors

26–27 January 2011
San Francisco, California, United States

Sponsored and Published by
IS&T—The Society for Imaging Science and Technology
SPIE

Volume 7874

The papers included in this volume were part of the technical conference cited on the cover and title page. Papers were selected and subject to review by the editors and conference program committee. Some conference presentations may not be available for publication. The papers published in these proceedings reflect the work and thoughts of the authors and are published herein as submitted. The publisher is not responsible for the validity of the information or for any outcomes resulting from reliance thereon.

Please use the following format to cite material from this book:

Author(s), "Title of Paper," in *Document Recognition and Retrieval XVIII*, edited by Gady Agam, Christian Viard-Gaudin, SPIE-IS&T Electronic Imaging, Vol. 7874 (SPIE, Bellingham, WA, 2011) Article CID Number.

ISSN 0277-786X
ISBN 9780819484116

Copublished by

SPIE

P.O. Box 10, Bellingham, Washington 98227-0010 USA
Telephone +1 360 676 3290 (Pacific Time) · Fax +1 360 647 1445
SPIE.org

and

IS&T—The Society for Imaging Science and Technology

7003 Kilworth Lane, Springfield, Virginia, 22151 USA
Telephone +1 703 642 9090 (Eastern Time) · Fax +1 703 642 9094
imaging.org

Copyright © 2011, Society of Photo-Optical Instrumentation Engineers

Copying of material in this book for internal or personal use, or for the internal or personal use of specific clients, beyond the fair use provisions granted by the U.S. Copyright Law is authorized by SPIE subject to payment of copying fees. The Transactional Reporting Service base fee for this volume is \$18.00 per article (or portion thereof), which should be paid directly to the Copyright Clearance Center (CCC), 222 Rosewood Drive, Danvers, MA 01923. Payment may also be made electronically through CCC Online at copyright.com. Other copying for republication, resale, advertising or promotion, or any form of systematic or multiple reproduction of any material in this book is prohibited except with permission in writing from the publisher. The CCC fee code is 0277-786X/11/\$18.00.

Printed in the United States of America.

Publication of record for individual papers is online in the SPIE Digital Library.

SPIE 
Digital Library

SPIDigitalLibrary.org

Paper Numbering: Proceedings of SPIE follow an e-First publication model, with papers published first online and then in print and on CD-ROM. Papers are published as they are submitted and meet publication criteria. A unique, consistent, permanent citation identifier (CID) number is assigned to each article at the time of the first publication. Utilization of CIDs allows articles to be fully citable as soon they are published online, and connects the same identifier to all online, print, and electronic versions of the publication. SPIE uses a six-digit CID article numbering system in which:

- The first four digits correspond to the SPIE volume number.
- The last two digits indicate publication order within the volume using a Base 36 numbering system employing both numerals and letters. These two-number sets start with 00, 01, 02, 03, 04, 05, 06, 07, 08, 09, 0A, 0B ... 0Z, followed by 10-1Z, 20-2Z, etc.

The CID number appears on each page of the manuscript. The complete citation is used on the first page, and an abbreviated version on subsequent pages. Numbers in the index correspond to the last two digits of the six-digit CID number.

Contents

vii *Conference Committee*

SESSION 1 INVITED PRESENTATION I

- 7874 02 **Scientific challenges underlying production document processing (Invited Paper)** [7874-01]
E. Saund, Palo Alto Research Ctr. (United States)

SESSION 2 CONTENT ANALYSIS

- 7874 03 **Automated identification of biomedical article type using support Vector machines** [7874-02]
I. C. Kim, D. X. Le, G. R. Thoma, National Library of Medicine (United States)
- 7874 04 **Introduction of statistical information in a syntactic analyzer for document image recognition** [7874-03]
A. O. Maroneze, B. Coüason, INSA (France); A. Lemaitre, Univ. of Rennes 2 (France)
- 7874 05 **High recall document content extraction** [7874-04]
C. An, H. S. Baird, Lehigh Univ. (United States)
- 7874 06 **Shape codebook based handwritten and machine printed text zone extraction** [7874-05]
J. Kumar, Univ. of Maryland, College Park (United States); R. Prasad, H. Cao, Raytheon BBN Technologies (United States); W. Abd-Almageed, D. Doermann, Univ. of Maryland, College Park (United States); P. Natarajan, Raytheon BBN Technologies (United States)

SESSION 3 RECOGNITION

- 7874 07 **A MRF model with parameter optimization by CRF for on-line recognition of handwritten Japanese characters** [7874-06]
B. Zhu, M. Nakagawa, Tokyo Univ. of Agriculture and Technology (Japan)
- 7874 08 **Improving a HMM-based off-line handwriting recognition system using MME-PSO optimization** [7874-07]
M. Hamdani, REGIM, Univ. of Sfax, ENIS (Tunisia); H. El Abed, Technische Univ. Braunschweig (Germany); T. M. Hamdani, REGIM, Univ. of Sfax, ENIS (Tunisia); V. Märgner, Technische Univ. Braunschweig (Germany); A. M. Alimi, REGIM, Univ. of Sfax, ENIS (Tunisia)
- 7874 09 **SemiBoost-based Arabic character recognition method** [7874-08]
B. Su, L. Peng, X. Ding, Tsinghua Univ. (China)

- 7874 0A **First experiments on a new online handwritten flowchart database** [7874-09]
A.-M. Awal, IRCCyN/IVC, CNRS, Ecole Polytechnique de l'Univ. De Nantes (France); G. Feng, Nanjing Univ. (China); H. Mouchère, C. Viard-Gaudin, IRCCyN/IVC, CNRS, Ecole Polytechnique de l'Univ. de Nantes (France)

SESSION 4 SEGMENTATION

- 7874 0B **Segmenting texts from outdoor images taken by mobile phones using color features** [7874-10]
Z. Liu, H. Zhou, Amazon.com (United States)
- 7874 0C **A perceptive method for handwritten text segmentation** [7874-11]
A. Lemaitre, Univ. de Rennes 2 (France); J. Camillerapp, B. Coüason, INSA (France)
- 7874 0D **Improved document image segmentation algorithm using multiresolution morphology** [7874-12]
S. S. Bukhari, Technical Univ. of Kaiserslautern (Germany); F. Shafait, German Research Ctr. for Artificial Intelligence (Germany); T. M. Breuel, Technical Univ. of Kaiserslautern (Germany)

SESSION 5 WRITER IDENTIFICATION OR VERIFICATION

- 7874 0F **Feature relevance analysis for writer identification** [7874-14]
I. Siddiqi, LIAPDE-SIP, Paris Descartes Univ. (France) and National Univ. of Sciences and Technology (Pakistan); K. Khurshid, LIAPDE-SIP, Paris Descartes Univ. (France) and Institute of Space Technology (Pakistan); N. Vincent, LIAPDE-SIP, Paris Descartes Univ. (France)
- 7874 0G **Using perturbed handwriting to support writer identification in the presence of severe data constraints** [7874-15]
J. Chen, W. Cheng, D. Lopresti, Lehigh Univ. (United States)
- 7874 0H **Statistical characterization of handwriting characteristics using automated tools** [7874-16]
G. R. Ball, S. N. Srihari, State Univ. of New York at Buffalo (United States)

SESSION 6 INFORMATION RETRIEVAL

- 7874 0I **Keyword and image-based retrieval of mathematical expressions** [7874-17]
R. Zanibbi, B. Yuan, Rochester Institute of Technology (United States)
- 7874 0J **Word spotting for handwritten documents using Chamfer Distance and Dynamic Time Warping** [7874-18]
R. M. Saabni, Ben-Gurion Univ. of the Negev (Israel) and Triangle Research and Development Ctr. (Israel); J. A. El-Sana, Triangle Research and Development Ctr. (Israel)
- 7874 0K **Automatic identification of ROI in figure images toward improving hybrid (text and image) biomedical document retrieval** [7874-19]
D. You, State Univ. of New York at Buffalo (United States); S. Antani, D. Demner-Fushman, M. M. Rahman, National Library of Medicine, National Institutes of Health (United States); V. Govindaraju, State Univ. of New York at Buffalo (United States); G. R. Thoma, National Library of Medicine, National Institutes of Health (United States)

- 7874 0L **Automatic extraction of numeric strings in unconstrained handwritten document images** [7874-20]
M. M. Haji, T. D. Bui, C. Y. Suen, Concordia Univ. (Canada)

SESSION 7 DOCUMENT RECOGNITION

- 7874 0M **Unsupervised method to generate page templates** [7874-21]
H. Déjean, Xerox Research Ctr. Europe (France)
- 7874 0N **Font group identification using reconstructed fonts** [7874-22]
M. P. Cutter, Univ. of Kaiserslautern (Germany); J. van Beusekom, F. Shafait, Univ. of Kaiserslautern (Germany) and German Research Ctr. for Artificial Intelligence (Germany); T. M. Breuel, Univ. of Kaiserslautern (Germany)
- 7874 0O **How carefully designed open resource sharing can help and expand document analysis research** [7874-23]
B. Lamiroy, Nancy Univ., LORIA (France); D. Lopresti, H. Korth, J. Heflin, Lehigh Univ. (United States)
- 7874 0P **Multiple-agent adaptation in whole-book recognition** [7874-24]
P. Xiu, H. S. Baird, Lehigh Univ. (United States)

SESSION 8 OCR ERROR AND BINARIZATION

- 7874 0Q **Ancient documents bleed-through evaluation and its application for predicting OCR error rates** [7874-25]
V. Rabeux, N. Journet, J. P. Domenger, LaBRI, Univ. de Bordeaux (France)
- 7874 0R **Binarization of camera-captured document using A MAP approach** [7874-26]
X. Peng, S. Setlur, V. Govindaraju, State Univ. of New York at Buffalo (United States); R. Sitaram, Hewlett-Packard Labs. India (India)
- 7874 0S **Statistical multi-resolution schemes for historical document binarization** [7874-27]
T. Obafemi-Ajayi, G. Agam, Illinois Institute of Technology (United States)

INTERACTIVE PAPER SESSION

- 7874 0T **A simple and effective figure caption detection system for old-style documents** [7874-28]
H. Zhou, Z. Liu, Amazon.com (United States)
- 7874 0U **Refining-driven paragraph recognition for electronic books in PDF** [7874-29]
J. Fang, Z. Tang, L. Gao, Peking Univ. (China)
- 7874 0V **Ruling line detection and removal** [7874-30]
E. Kavallieratou, Univ. of the Aegean (Greece); D. Lopresti, J. Chen, Lehigh Univ. (United States)

- 7874 0W **Natural scene logo recognition by joint boosting feature selection in salient regions** [7874-31]
W. Fan, J. Sun, S. Naoi, Fujitsu Research and Development Ctr. Co., Ltd. (China);
A. Minagawa, Y. Hotta, Fujitsu Labs., Ltd. (Japan)
- 7874 0X **A framework to improve digital corpus uses: image-mode navigation** [7874-32]
L. Eynard, CNRS, INSA-Lyon, LIRIS, Univ. Lyon (France); V. Malleron, CNRS, INSA-Lyon, LIRIS,
Univ. Lyon (France) and CNRS, LIRE, Univ. Lyon 2 (France); H. Emptoz, CNRS, INSA-Lyon, LIRIS,
Univ. Lyon (France)
- 7874 0Y **Parameter calibration for synthesizing realistic-looking variability in offline handwriting**
[7874-33]
W. Cheng, D. Lopresti, Lehigh Univ. (United States)
- 7874 0Z **Automatic segmentation of subfigure image panels for multimodal biomedical document
retrieval** [7874-34]
B. Cheng, Missouri Univ. of Science and Technology (United States); S. Antani, National
Library of Medicine, National Institutes of Health (United States); R. J. Stanley, Missouri Univ.
of Science and Technology (United States); G. R. Thoma, National Library of Medicine,
National Institutes of Health (United States)
- 7874 10 **A new method for perspective correction of document images** [7874-35]
J. Rodríguez-Piñeiro, Univ. of Vigo (Spain); P. Comesaña-Alfaro, Univ. of Vigo (Spain) and
Univ. of New Mexico (United States); F. Pérez-González, Univ. of Vigo (Spain), Univ. of New
Mexico (United States), and Gradient (Spain); A. Malvido-García, Bit Oceans Research
(Spain)
- 7874 11 **Robust keyword retrieval method for OCRed text** [7874-36]
Y. Fujii, H. Takebe, H. Tanaka, Y. Hotta, Fujitsu Labs., Ltd. (Japan)
- 7874 12 **Online medical symbol recognition using a Tablet PC** [7874-37]
A. Kundu, Q. Hu, S. Boykin, C. Clark, R. Fish, S. Jones, S. Moore, MITRE Corp. (United States)
- 7874 13 **Characterizing challenged Minnesota ballots** [7874-38]
G. Nagy, Rensselaer Polytechnic Institute (United States); D. Lopresti, Lehigh Univ. (United
States); E. H. Barney Smith, Boise State Univ. (United States); Z. Wu, Rensselaer Polytechnic
Institute (United States)
- 7874 14 **A mask-based enhancement method for historical documents** [7874-39]
E. H. Barney Smith, Boise State Univ. (United States); J. Darbon, CMLA, ENS Cachan, CNRS,
PRES Univ. (France); L. Likforman-Sulem, Telecom ParisTech (France)
- 7874 15 **Document image retrieval with morphology-based segmentation and features
combination** [7874-40]
T. C. Bockholt, G. D. C. Cavalcanti, C. A. B. Mello, Federal Univ. of Pernambuco (Brazil)
- 7874 16 **Boosting based text and non-text region classification** [7874-41]
B. Xie, G. Agam, Illinois Institute of Technology (United States)
- 7874 17 **OMR of early plainchant manuscripts in square notation: a two-stage system** [7874-42]
C. Ramirez, J. Ohya, Waseda Univ. (Japan)

Author Index

Conference Committee

Symposium Chair

Sabine E. Süsstrunk, Ecole Polytechnique Fédérale de Lausanne
(Switzerland)

Symposium Cochair

Majid Rabbani, Eastman Kodak Company (United States)

Conference Chairs

Gady Agam, Illinois Institute of Technology (United States)
Christian Viard-Gaudin, Université de Nantes (France)

Program Committee

Apostolos Antonacopoulos, University of Salford (United Kingdom)
Elisa H. Barney Smith, Boise State University (United States)
Kathrin Berkner, Ricoh Innovations, Inc. (United States)
Xiaoqing Ding, Tsinghua University (China)
David S. Doermann, University of Maryland, College Park (United States)
Oleg D. Golubitsky, Google, Inc. (Canada)
Jiaying Hu, IBM Thomas J. Watson Research Center (United States)
Laurence Likforman-Sulem, Telecom ParisTech (France)
Marcus Liwicki, Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (Germany)
Xiaofan Lin, Vobile, Inc. (United States)
Daniel P. Lopresti, Lehigh University (United States)
Hiroshi Sako, Hitachi, Ltd. (Japan)
Lambert R. B. Schomaker, University of Groningen (Netherlands)
Sargur N. Srihari, University at Buffalo (United States)
Venkata Subramaniam, IBM India Research Laboratory (India)
Kazem Taghva, University of Nevada, Las Vegas (United States)
George R. Thoma, National Library of Medicine (United States)
Alessandro Vinciarelli, University of Glasgow (United Kingdom)
Berrin Yanikoglu, Sabanci University (Turkey)
Jie Zou, National Library of Medicine (United States)

