

Document Recognition and Retrieval XIX

**Christian Viard-Gaudin
Richard Zanibbi**
Editors

**25–26 January 2012
Burlingame, California, United States**

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

Volume 8297

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 publishers are 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 XIX*, edited by Christian Viard-Gaudin, Richard Zanibbi, Proceedings of SPIE-IS&T Electronic Imaging, SPIE Vol. 8297, Article CID Number (2012).

ISSN 0277-786X
ISBN 9780819489449

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 © 2012, Society of Photo-Optical Instrumentation Engineers and The Society for Imaging Science and Technology.

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 the publishers 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/12/\$18.00.

Printed in the United States of America.

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 as 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	<i>Conference Committee</i>
ix	<i>Introduction</i>

REGION LABELING

- 8297 03 **Graphical image classification combining an evolutionary algorithm and binary particle swarm optimization** [8297-02]
B. Cheng, R. Wang, Missouri Univ. of Science and Technology (United States); S. Antani, National Library of Medicine (United States); R. J. Stanley, Missouri Univ. of Science and Technology (United States); G. R. Thoma, National Library of Medicine (United States)
- 8297 04 **Combining SVM classifiers to identify investigator name zones in biomedical articles** [8297-03]
J. Kim, D. X. Le, G. R. Thoma, National Library of Medicine (United States)
- 8297 05 **Comprehensive color segmentation system for noisy digitized documents to enhance text extraction** [8297-04]
A. Ouji, Y. Leydier, F. LeBourgeois, Institut National des Sciences Appliquées, CNRS, Univ. de Lyon (France)
- 8297 06 **Ensemble methods with simple features for document zone classification** [8297-05]
T. Obafemi-Ajayi, Univ. of Missouri (United States); G. Agam, B. Xie, Illinois Institute of Technology (United States)

HANDWRITING RECOGNITION

- 8297 07 **A robust omnifont open-vocabulary Arabic OCR system using pseudo-2D-HMM** [8297-06]
A. M. Rashwan, M. A. Rashwan, A. Abdel-Hameed, S. Abdou, Cairo Univ. (Egypt) and RDI (Egypt); A. H. Khalil, Cairo Univ. (Egypt)
- 8297 08 **Variable length and context-dependent HMM letter form models for Arabic handwritten word recognition** [8297-07]
A.-L. Bianne-Bernard, A2iA SA (France) and Lab. Traitement et Communication de l'Information, CNRS, Telecom ParisTech (France); F. Menasri, A2iA SA (France); L. Likforman-Sulem, Lab. Traitement et Communication de l'Information, CNRS, Telecom ParisTech (France); C. Mokbel, Univ. of Balamand (Lebanon); C. Kermorvant, A2iA SA (France)
- 8297 09 **Post processing for offline Chinese handwritten character string recognition** [8297-08]
Y. Wang, X. Ding, C. Liu, Tsinghua Univ. (China)
- 8297 0A **Complexity reduction with recognition rate maintained for online handwritten Japanese text recognition** [8297-09]
J. Gao, B. Zhu, M. Nakagawa, Tokyo Univ. of Agriculture and Technology (Japan)

- 8297 0B **Improving isolated and in-context classification of handwritten characters** [8297-10]
V. Mazalov, S. M. Watt, The Univ. of Western Ontario (Canada)

GRAPHICS RECOGNITION

- 8297 0C **Using specific evaluation for comparing and combining competing algorithms: applying it to table column detection** [8297-11]
A. C. e Silva, LIAAD-INESC Porto (Portugal)
- 8297 0D **Identification of embedded mathematical formulas in PDF documents using SVM** [8297-12]
X. Lin, L. Gao, Peking Univ. (China); Z. Tang, Peking Univ. (China) and State Key Lab. of Digital Publishing Technology (China); X. Hu, BeiHang Univ. (China); X. Lin, Vobile, Inc. (United States)
- 8297 0E **Chemical structure recognition: a rule-based approach** [8297-13]
N. M. Sadawi, A. P. Sexton, V. Sorge, The Univ. of Birmingham (United Kingdom)
- 8297 0F **Quantify spatial relations to discover handwritten graphical symbols** [8297-14]
J. Li, H. Mouchère, C. Viard-Gaudin, Institut de Recherche en Communications et Cybernétique, CNRS, Univ. de Nantes (France)

INFORMATION RETRIEVAL

- 8297 0H **Automatic indexing of scanned documents: a layout-based approach** [8297-16]
D. Esser, D. Schuster, K. Muthmann, Technische Univ. Dresden (Germany); M. Berger, DocuWare AG (Germany); A. Schill, Technische Univ. Dresden (Germany)
- 8297 0I **Layout-based substitution tree indexing and retrieval for mathematical expressions** [8297-17]
T. Schellenberg, B. Yuan, R. Zanibbi, Rochester Institute of Technology (United States)

HUMAN COMPUTER INTERACTION

- 8297 0J **Efficient cost-sensitive human-machine collaboration for offline signature verification** [8297-18]
J. Coetzer, J. Swanepoel, Stellenbosch Univ. (South Africa); R. Sabourin, Univ. of Quebec (Canada)
- 8297 0K **Questioned document workflow for handwriting with automated tools** [8297-19]
K. Das, S. N. Srihari, Univ. at Buffalo (United States); H. Srinivasan, Janya Inc. (United States)
- 8297 0L **Iterative analysis of document collections enables efficient human-initiated interaction** [8297-20]
J. Chazalon, B. Coüasnon, Institut National des Sciences Appliquées de Rennes, CNRS, Univ. Européenne de Bretagne (France)

8297 OM **VeriClick: an efficient tool for table format verification** [8297-21]
G. Nagy, M. Tamhankar, Rensselaer Polytechnic Institute (United States)

8297 ON **Asymptotic cost in document conversion** [8297-22]
D. Blostein, Queen's Univ. (Canada); G. Nagy, Rensselaer Polytechnic Institute (United States)

STYLE OR WRITER IDENTIFICATION

8297 OO **Style comparisons in calligraphy** [8297-23]
X. Zhang, Shanghai Maritime Univ. (China); G. Nagy, Rensselaer Polytechnic Institute (United States)

8297 OP **An Oracle-based co-training framework for writer identification in offline handwriting** [8297-24]
U. Porwal, Univ. at Buffalo (United States); S. Rajan, Fujitsu Labs. of America, Inc. (United States); V. Govindaraju, Univ. at Buffalo (United States)

8297 OQ **Handwritten document age classification based on handwriting styles** [8297-25]
C. Ramaiah, G. Kumar, V. Govindaraju, Univ. at Buffalo (United States)

8297 OR **Handwriting individualization using distance and rarity** [8297-26]
Y. Tang, S. Srihari, Univ. at Buffalo (United States); H. Srinivasan, Janya Inc. (United States)

8297 OS **Construction of language models for an handwritten mail reading system** [8297-27]
O. Morillot, L. Likforman-Sulem, Lab. Traitement et Communication de l'Information, CNRS, Telecom ParisTech (France); E. Grosicki, DGA (France)

INTERACTIVE PAPER SESSION

8297 OT **Bleed-through removal in degraded documents** [8297-28]
R. Rowley-Brooke, A. Kokaram, Trinity College Dublin (Ireland)

8297 OU **Clustering document fragments using background color and texture information** [8297-29]
S. Chanda, K. Franke, Gjøvik Univ. College (Norway); U. Pal, Indian Statistical Institute (India)

8297 OV **Lecture video segmentation and indexing** [8297-30]
D. Ma, G. Agam, Illinois Institute of Technology (United States)

8297 OW **Unsupervised categorization method of graphemes on handwritten manuscripts: application to style recognition** [8297-31]
H. Daher, D. Gaceb, V. Eglin, S. Bres, Institut National des Sciences Appliquées, CNRS, Univ. de Lyon (France); N. Vincent, Lab. d'Informatique Paris Descartes, CNRS, Univ. Paris Descartes (France)

8297 OX **Retrieving handwriting by combining word spotting and manifold ranking** [8297-32]
S. Peña Saldarriaga, Synchromedia (Canada); E. Morin, Lab. Informatique de Nantes Atlantique, CNRS, Univ. de Nantes (France); C. Viard-Gaudin, Institut de Recherche en Communications et Cybernétique de Nantes, CNRS, Univ. de Nantes (France)

- 8297 0Y **The A2iA French handwriting recognition system at the Rimes-ICDAR2011 competition**
[8297-33]
F. Menasri, J. Louradour, A.-L. Bianne-Bernard, C. Kermorvant, A2iA SA (France)
- 8297 0Z **Using connected component decomposition to detect straight line segments in documents**
[8297-34]
X. Feng, A. Youssef, The George Washington Univ. (United States)
- 8297 10 **A synthetic document image dataset for developing and evaluating historical document processing methods** [8297-35]
D. Walker, W. Lund, E. Ringger, Brigham Young Univ. (United States)

Author Index

Conference Committee

Symposium Chairs

Majid Rabbani, Eastman Kodak Company (United States)
Gaurav Sharma, University of Rochester (United States)

Conference Chairs

Christian Viard-Gaudin, Université de Nantes (France)
Richard Zanibbi, Rochester Institute of Technology (United States)

Program Committee

Gady Agam, Illinois Institute of Technology (United States)
Elisa H. Barney Smith, Boise State University (United States)
Bill Barrett, Brigham Young University (United States)
Kathrin Berkner, Ricoh Innovations, Inc. (United States)
Bertrand Coüasnon, Institut National des Sciences Appliquées de
Rennes (France)
Hervé Déjean, Xerox Research Centre Europe Grenoble (France)
Xiaoqing Ding, Tsinghua University (China)
David Scott Doermann, University of Maryland, College Park (United
States)
Oleg D. Golubitsky, Google Waterloo (Canada)
Jianying Hu, IBM Thomas J. Watson Research Center (United States)
Laurence Likforman-Sulem, Telecom ParisTech (France)
Xiaofan Lin, Vobile, Inc. (United States)
Marcus Liwicki, Deutsches Forschungszentrum für Künstliche Intelligenz
GmbH (Germany)
Daniel P. Lopresti, Lehigh University (United States)
Hiroshi Sako, Hosei University (Japan)
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)
Berrin Yanikoglu, Sabanci University (Turkey)
Jie Zou, National Library of Medicine (United States)

Introduction

On behalf of the DRR 2012 program committee, we are honored and delighted to welcome you to San Francisco, CA, USA for the Nineteenth Document Recognition and Retrieval Conference (DRR 2012). The conference is part of the Electronic Imaging Symposium, which brings together researchers from various backgrounds related to electronic imaging for an exciting research event.

As one of the well-established conferences of the Electronic Imaging Symposium organized by The Society for Imaging Science and Technology (IS&T), and the International Association for Optics and Photonics (SPIE), DRR continues to reflect the vitality and impact of the worldwide document analysis research community. This year the conference received 46 submissions, fewer than the record 54 received for DRR 2011. However, the quality of the papers was quite high, leading to roughly 59% being accepted for oral presentation, and an additional 17% for poster presentation. We proudly invite you to take advantage of the excellent technical program for DRR 2012, which includes recent advances in document image processing and layout analysis, handwriting recognition, graphics recognition (e.g. for tables, chemical diagrams, and math), human factors in system training and deployment, and retrieval using document images and graphics as queries.

As the Program Chairs, we take primary responsibility for deciding the program, but its content and quality derive from the vision and hard work of the 23 members of the program committee, who have our sincerest thanks for providing paper reviews within a very short time during the summer. Their reviews and comments were essential to shaping the technical program.

We also extend our warmest thanks to Dr. Samy Bengio (Google) and Dr. Christopher Manning (Stanford University), who will give keynote talks on "Large scale visual semantic extraction" (Dr. Bengio) and "Language modeling for Information Retrieval" (Dr. Manning). Their talks will contribute greatly to the success of the conference.

We will announce the winner of the best student paper during the conference. The awardee will be presented with a certificate and monetary award check at the Plenary Session and Society Award Presentations.

We wish you all an exciting and fruitful stay in San Francisco!

Christian Viard-Gaudin
Richard Zanibbi

