

Document Recognition and Retrieval XVII

Laurence Likforman-Sulem
Gady Agam
Editors

19–21 January 2010
San Jose, California, United States

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

Cosponsored by
Institut TELECOM (France)

Volume 7534

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 XVII*, edited by Laurence Likforman-Sulem, Gady Agam, Proceedings of SPIE-IS&T Electronic Imaging, SPIE Vol. 7534, Article CID Number (2010).

ISSN 0277-786X

ISBN 9780819479273

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 © 2010, 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/10/\$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 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> |

SESSION 1 INVITED PRESENTATION I

- 7534 02 **A general approach to discovering, registering, and extracting features from raster maps (Invited Paper)** [7534-01]
C. A. Knoblock, Univ. of Southern California (United States) and Geosemble Technologies (United States); C.-C. Chen, Geosemble Technologies (United States); Y.-Y. Chiang, A. Goel, Univ. of Southern California (United States); M. Michelson, Fetch Technologies (United States); C. Shahabi, Univ. of Southern California (United States) and Geosemble Technologies (United States)

SESSION 2 INFORMATION RETRIEVAL

- 7534 03 **Combining approaches to on-line handwriting information retrieval** [7534-02]
S. Peña Saldarriaga, LINA, CNRS, Univ. de Nantes (France); C. Viard-Gaudin, IRCCyN, CNRS, École Polytechnique de l'Univ. de Nantes (France); E. Morin, LINA, CNRS, Univ. de Nantes (France)
- 7534 04 **A stacked sequential learning method for investigator name recognition from web-based medical articles** [7534-03]
X. Zhang, J. Zou, D. X. Le, G. Thoma, National Library of Medicine (United States)
- 7534 05 **Numbered sequence detection in documents** [7534-04]
H. Déjean, Xerox Research Ctr. Europe (France)
- 7534 06 **Date of birth extraction using precise shallow parsing** [7534-05]
R. Pereda, K. Taghva, Univ. of Nevada, Las Vegas (United States)

SESSION 3 CONTENT ANALYSIS

- 7534 07 **The aware toolbox for the detection of law infringements on web pages** [7534-06]
A. Shahab, T. Kieninger, A. Dengel, German Research Ctr. for Artificial Intelligence GmbH (Germany)
- 7534 08 **On the usability and security of pseudo-signatures** [7534-07]
J. Chen, D. Lopresti, Lehigh Univ. (United States)
- 7534 09 **Time and space optimization of document content classifiers** [7534-08]
D. Yin, H. S. Baird, C. An, Lehigh Univ. (United States)

- 7534 0A **Detecting modifications in paper documents: a coding approach** [7534-09]
Y. Sankarasubramaniam, B. Narayanan, K. Viswanathan, A. Kuchibhotla, Hewlett-Packard Labs. India (India)

SESSION 4 TEXT LINE AND SEGMENTATION

- 7534 0B **General text line extraction approach based on locally orientation estimation** [7534-10]
N. Ouwayed, A. Belaïd, LORIA, Univ. of Nancy 2 (France); F. Auger, Univ. of Nantes, IREENA (France)
- 7534 0C **Semi-supervised learning for detecting text-lines in noisy document images** [7534-11]
Z. Liu, H. Zhou, Amazon.com (United States)
- 7534 0D **Touching character segmentation method for Chinese historical documents** [7534-12]
X. Sun, L. Peng, X. Ding, Tsinghua Univ. (China)

SESSION 5 INVITED PRESENTATION II

- 7534 0E **Technologies for developing an advanced intelligent ATM with self-defence capabilities (Invited Paper)** [7534-13]
H. Sako, Hitachi, Ltd. (Japan)

SESSION 6 DOCUMENT IMAGE PROCESSING

- 7534 0F **Learning shape features for document enhancement** [7534-14]
T. Obafemi-Ajayi, G. Agam, O. Frieder, Illinois Institute of Technology (United States)
- 7534 0G **Enhancement of camera-based whiteboard images** [7534-15]
Y. He, J. Sun, S. Naoi, Fujitsu Research and Development Ctr. Co., Ltd. (China);
A. Minagawa, Y. Hotta, Fujitsu Labs., Ltd. (Japan)
- 7534 0H **Effect of pre-processing on binarization** [7534-16]
E. H. Barney Smith, Boise State Univ. (United States); L. Likforman-Sulem, Telecom ParisTech (France); J. Darbon, Univ. of California, Los Angeles (United States)

SESSION 7 RECOGNITION I

- 7534 0I **Context-dependent HMM modeling using tree-based clustering for the recognition of handwritten words** [7534-17]
A.-L. Bianne, A2iA SA (France) and Telecom ParisTech/TSI, CNRS, LTCI (France);
C. Kermorvant, A2iA SA (France); L. Likforman-Sulem, Telecom ParisTech/TSI, CNRS, LTCI (France)
- 7534 0J **Font adaptation of an HMM-based OCR system** [7534-18]
K. Ait-Mohand, L. Heutte, T. Paquet, Univ. de Rouen (France); N. Ragot, Univ. François Rabelais Tours (France)

- 7534 OK **A new pre-classification method based on associative matching method** [7534-19]
Y. Katsuyama, A. Minagawa, Y. Hotta, Fujitsu Labs. Ltd. (Japan); S. Omachi, N. Kato, Tohoku Univ. (Japan)
- 7534 OL **A neural-linguistic approach for the recognition of a wide Arabic word lexicon** [7534-20]
I. Ben Cheikh, A. Kacem, UTIC-ESSTT (Tunisia); A. Belaid, LORIA (France)

SESSION 8 RECOGNITION II

- 7534 OM **Incorporating linguistic post-processing into whole-book recognition** [7534-21]
P. Xiu, H. S. Baird, Lehigh Univ. (United States)
- 7534 ON **A word language model based contextual language processing on Chinese character recognition** [7534-22]
C. Huang, X. Ding, Y. Chen, Tsinghua Univ. (China)
- 7534 OO **Efficient automatic OCR word validation using word partial format derivation and language model** [7534-23]
S. Chen, D. Misra, G. R. Thoma, U.S. National Library of Medicine (United States)
- 7534 OP **Comparison of historical documents for writership** [7534-24]
G. R. Ball, D. Pu, CEDAR, Univ. at Buffalo (United States); R. Stritmatter, Coppin State Univ. (United States); S. N. Srihari, CEDAR, Univ. at Buffalo (United States)

SESSION 9 DOCUMENT STRUCTURE RECOGNITION

- 7534 OQ **Interactive-predictive detection of handwritten text blocks** [7534-25]
O. Ramos Terrades, N. Serrano, Univ. Politècnica de València (Spain); A. Gordó, E. Valveny, Univ. Autònoma de Barcelona (Spain); A. Juan, Univ. Politècnica de València (Spain)
- 7534 OR **Using definite clause grammars to build a global system for analyzing collections of documents** [7534-26]
J. Chazalon, B. Coüasnon, Univ. Européenne de Bretagne (France)
- 7534 OS **Detection of figure and caption pairs based on disorder measurements** [7534-27]
C. Faure, CNRS-LTCl, TELECOM-ParisTech (France); N. Vincent, LIPADE, Univ. Paris Descartes (France)

INTERACTIVE PAPER SESSION

- 7534 OT **Evaluation of human perception of degradation in document images** [7534-28]
T. Obafemi-Ajayi, G. Agam, O. Frieder, Illinois Institute of Technology (United States)
- 7534 OU **Naïve Bayes and SVM classifiers for classifying databank accession number sentences from online biomedical articles** [7534-29]
J. Kim, D. X. Le, G. R. Thoma, National Library of Medicine (United States)

- 7534 0V **Biomedical article retrieval using multimodal features and image annotations in region-based CBIR** [7534-30]
D. You, SUNY at Buffalo (United States); S. Antani, D. Demner-Fushman, M. M. Rahman, National Library of Medicine (United States); V. Govindaraju, SUNY at Buffalo (United States); G. R. Thoma, National Library of Medicine (United States)
- 7534 0W **Trainable multiscript orientation detection** [7534-31]
J. van Beusekom, German Research Ctr. for Artificial Intelligence GmbH (Germany) and Technical Univ. of Kaiserslautern (Germany); Y. Rangoni, German Research Ctr. for Artificial Intelligence GmbH (Germany); T. M. Breuel, German Research Ctr. for Artificial Intelligence GmbH (Germany) and Technical Univ. of Kaiserslautern (Germany)
- 7534 0X **Improved CHAID algorithm for document structure modelling** [7534-32]
A. Belaïd, T. Moinel, Y. Rangoni, LORIA, Univ. Nancy 2 (France)
- 7534 0Y **Ant colony optimization with selective evaluation for feature selection in character recognition** [7534-33]
I.-S. Oh, Chonbuk National Univ. (Korea, Republic of); J.-S. Lee, Woosuk Univ. (Korea, Republic of)
- 7534 0Z **Analysis of line structure in handwritten documents using the Hough transform** [7534-34]
G. R. Ball, H. Kasiviswanathan, S. N. Srihari, A. Narayanan, CEDAR, Univ. at Buffalo (United States)
- 7534 10 **A hybrid classifier for handwritten mathematical expression recognition** [7534-35]
A.-M. Awal, H. Mouchère, C. Viard-Gaudin, IRCCyN/IVC, CNRS, Ecole Polytechnique de l'Univ. de Nantes (France)
- 7534 11 **A combined recognition system for online handwritten Pinyin input** [7534-36]
M. Zhu, C. Liu, Tsinghua Univ. (China)

Author Index

Conference Committee

Symposium Chair

Jan P. Allebach, Purdue University (United States)

Symposium Cochair

Sabine Süsstrunk, Ecole Polytechnique Fédérale de Lausanne
(Switzerland)

Conference Chairs

Laurence Likforman-Sulem, Telecom ParisTech (France)
Gady Agam, Illinois Institute of Technology (United States)

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, The University of Western Ontario (Canada)
Jianying Hu, IBM Thomas J. Watson Research Center (United States)
Xiaofan Lin, Vobile, Inc. (United States)
Marcus Liwicki, Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (Germany)
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)
Christian Viard-Gaudin, Université de Nantes (France)
Alessandro Vinciarelli, Idiap Research Institute (Switzerland)
Berrin Yanikoglu, Sabancı University (Turkey)
Jie Zou, National Library of Medicine (United States)

Introduction

This volume brings together the papers presented at the Document Recognition and Retrieval XVII (DRR) conference held January 2010 in San Jose as part of the IS&T/SPIE Electronic Imaging symposium. The DRR conference aims at presenting approaches which may improve the extraction of information from various types of documents: scanned, web-based, on-line, or camera-based documents. These approaches may concern the early stages of information extraction such as image enhancement, binarization, or document structure extraction. They may also concern higher stages of information extraction such as OCR, handwriting recognition, or symbol recognition. When these stages are achieved or when an electronic version is already available, information extraction and retrieval techniques can be applied. Related studies include document or writer/signature authentication. The 36 articles included in this volume deal with those topics for documents as various as historical documents, handwritten mails, maps or medical articles, and for documents written in diverse languages.

Our two invited speakers this year are Craig Knoblock from the University of Southern California and Hiroshi Sako from Hitachi Japan. Craig Knoblock will give a talk on the processing of maps and map discovery. Hiroshi Sako will speak about advanced automated teller machines (ATMs) which include intelligence for preventing different types of attacks.

This year we continue the tradition of giving the Best Student Paper Award to a paper whose lead author is a full-time student. We gratefully acknowledge Institut TELECOM and Telecom ParisTech for sponsoring this award.

We wish to thank the program committee of DRR for providing the three required reviews and the additional reviewers for assisting in the review process. Many thanks go to the SPIE conference organizers for their help in organizing the conference. At last, we thank all the participating authors of the papers for their contributions to this conference and we welcome any feedback and suggestions.

We encourage you to actively use this conference as an opportunity for discussions with your colleagues in this unique interdisciplinary framework, and we wish you a very fruitful stay in San Jose and California.

**Laurence Likforman-Sulem
Gady Agam**

