PROCEEDINGS OF SPIE

SPIDigitalLibrary.org/conference-proceedings-of-spie

Front Matter: Volume 8658

, "Front Matter: Volume 8658," Proc. SPIE 8658, Document Recognition and Retrieval XX, 865801 (5 February 2013); doi: 10.1117/12.2020094

SPIE.

Event: IS&T/SPIE Electronic Imaging, 2013, Burlingame, California, United States

Document Recognition and Retrieval XX

**Richard Zanibbi
Bertrand Couasnon**
Editors

**5–7 February 2013
Burlingame, California, United States**

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

Cosponsored by
Qualcomm Inc. (United States)
Google Inc. (United States)

Published by
SPIE

Volume 8658

Proceedings of SPIE 0277-786X, V. 8658

Document Recognition and Retrieval XX, edited by Richard Zanibbi, Bertrand Couasnon,
Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 8658, 865801 · © 2013 SPIE-IS&T
CCC code: 0277-786/13/\$18 · doi: 10.1117/12.2020094

SPIE-IS&T/ Vol. 8658 865801-1

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 XX*, edited by Richard Zanibbi, Bertrand Coùasnon, Proceedings of SPIE-IS&T Electronic Imaging, SPIE Vol. 8658. Article CID Number (2013)

ISSN: 0277-786X

ISBN: 9780819494313

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 © 2013, 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/13/\$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

ix	Conference Committee
xi	Introduction

KEYNOTE SESSION

8658 02	History of the Tesseract OCR engine: what worked and what didn't (Keynote Paper) [8658-15] R. Smith, Google Inc. (United States)
---------	---

SESSION 1 IMAGE-BASED RETRIEVAL

8658 04	Semi-structured document image matching and recognition [8658-1] O. Augereau, N. Journet, J.-P. Domenger, Univ. de Bordeaux (France)
8658 05	Rotation-robust math symbol recognition and retrieval using outer contours and image subsampling [8658-2] S. Zhu, L. Hu, R. Zanibbi, Rochester Institute of Technology (United States)
8658 06	NESP: Nonlinear enhancement and selection of plane for optimal segmentation and recognition of scene word images [8658-3] D. Kumar, M. N. Anil Prasad, A. G. Ramakrishnan, Indian Institute of Science (India)

SESSION 2 HANDWRITING

8658 07	Combining evidence using likelihood ratios in writer verification [8658-4] S. Srihari, D. Kovalenko, Y. Tang, G. Ball, Univ. at Buffalo, SUNY (United States)
8658 08	Handwritten word preprocessing for database adaptation [8658-5] C. Oprean, L. Likforman-Sulem, Institut Mines-Telecom, CNRS, Telecom ParisTech (France); C. Mokbel, Univ. of Balamand (Lebanon)
8658 09	Optimal policy for labeling training samples [8658-6] L. Lipsky, Univ. of Connecticut (United States); D. Lopresti, Lehigh Univ. (United States); G. Nagy, Rensselaer Polytechnic Institute (United States)
8658 0A	Evaluation of lexicon size variations on a verification and rejection system based on SVM, for accurate and robust recognition of handwritten words [8658-7] Y. Ricquebourg, B. Coüasnon, IRISA / INSA, Univ. Européenne de Bretagne (France); L. Guichard, E2I SAS (France)

SESSION 3 LAYOUT ANALYSIS

- 8658 0B **Comic image understanding based on polygon detection** [8658-8]
L. Li, Y. Wang, Z. Tang, D. Liu, Peking Univ. (China)
- 8658 0C **Context modeling for text/non-text separation in free-form online handwritten documents**
[8658-9]
A. Delaye, C.-L. Liu, Institute of Automation (China)
- 8658 0D **Annotating image ROIs with text descriptions for multimodal biomedical document
retrieval** [8658-10]
D. You, M. Simpson, S. Antani, D. Demner-Fushman, G. R. Thoma, National Library of
Medicine (United States)
- 8658 0E **Graphic composite segmentation for PDF documents with complex layouts** [8658-11]
C. Xu, Z. Tang, Peking Univ. (China) and Zhongguancun Haidian Science Park (China);
X. Tao, C. Shi, Peking Univ. (China)

SESSION 4 WORD AND SYMBOL SPOTTING

- 8658 0F **A classification-free word-spotting system** [8658-12]
N. Vasilopoulos, E. Kavallieratou, Univ. of the Aegean (Greece)
- 8658 0G **Combining geometric matching with SVM to improve symbol spotting** [8658-13]
N. Nayef, T. M. Breuel, Technical Univ. of Kaiserslautern (Germany)
- 8658 0H **Segmentation-free keyword spotting framework using dynamic background model**
[8658-14]
G. Kumar, S. Wshah, V. Govindaraju, Univ. at Buffalo, SUNY (United States);
S. Ramachandula, Hewlett-Packard Labs. (India)

SESSION 5 HISTORICAL DOCUMENTS

- 8658 0I **Data acquisition from cemetery headstones** [8658-16]
C. S. Christiansen, W. A. Barrett, Brigham Young Univ. (United States)
- 8658 0J **Automated recognition and extraction of tabular fields for the indexing of census records**
[8658-17]
R. Clawson, K. Bauer, G. Chidester, M. Pohontsch, D. Kennard, J. Ryu, W. Barrett, Brigham
Young Univ. (United States)
- 8658 0K **Old document image segmentation using the autocorrelation function and multiresolution
analysis** [8658-18]
M. Mehri, L3i, Univ. of La Rochelle (France) and LITIS, Univ. of Rouen (France);
P. Gomez-Krämer, L3i, Univ. of La Rochelle (France); P. Héroux, LITIS, Univ. of Rouen
(France); R. Mullot, L3i, Univ. of La Rochelle (France)
- 8658 0L **Lexicon-supported OCR of eighteenth century Dutch books: a case study** [8658-19]
J. de Does, K. Depuydt, INL (Netherlands)

SESSION 6 ARABIC AND CHINESE CHARACTER RECOGNITION

- 8658 OM **Character feature integration of Chinese calligraphy and font** [8658-20]
C. Shi, J. Xiao, W. Jia, Peking Univ. (China); C. Xu, Peking Univ. (China) and Zhongguancun Haidian Science Park (China)
- 8658 ON **A segmentation-free approach to Arabic and Urdu OCR** [8658-21]
N. Sabbour, German Univ. in Cairo (Egypt); F. Shafait, German Research Ctr. for Artificial Intelligence (Germany)
- 8658 OO **Local projection-based character segmentation method for historical Chinese documents** [8658-22]
L. Yang, L. Peng, Tsinghua Univ. (China)

SESSION 7 INTERACTIVE PAPER SESSION

- 8658 OP **A super resolution framework for low resolution document image OCR** [8658-34]
D. Ma, G. Agam, Illinois Institute of Technology (United States)
- 8658 OQ **A robust pointer segmentation in biomedical images toward building a visual ontology for biomedical article retrieval** [8658-35]
D. You, M. Simpson, S. Antani, D. Demner-Fushman, G. R. Thoma, National Library of Medicine (United States)
- 8658 OR **Combining multiple thresholding binarization values to improve OCR output** [8658-36]
W. B. Lund, D. J. Kennard, E. K. Ringger, Brigham Young Univ. (United States)
- 8658 OS **Goal-oriented evaluation of binarization algorithms for historical document images** [8658-37]
T. Obafemi-Ajayi, Univ. of Missouri-Columbia (United States); G. Agam, Illinois Institute of Technology (United States)
- 8658 OT **Document segmentation via oblique cuts** [8658-38]
J. Svendsen, A. Branzan-Albu, Univ. of Victoria (Canada)
- 8658 OU **Preprocessing document images by resampling is error prone and unnecessary** [8658-39]
G. Nagy, Rensselaer Polytechnic Institute (United States)
- 8658 OV **Multilingual artificial text detection and extraction from still images** [8658-40]
A. Raza, A. Abidi, National Univ. of Sciences and Technology (Pakistan); I. Siddiqi, Bahria Univ. (Pakistan)
- 8658 OW **A proposal system for historic Arabic manuscript transcription and retrieval** [8658-41]
A. Labben, A. Kacem, LaTICE (Tunisia); A. Belaïd, LORIA (France)
- 8658 OX **Evaluation of document binarization using eigen value decomposition** [8658-42]
D. Kumar, M. N. Anil Prasad, A. G. Ramakrishnan, Indian Institute of Science (India)

- 8658 0Y **Efficient symbol retrieval by building a symbol index from a collection of line drawings** [8658-43]
N. Nayef, T. M. Breuel, Technical Univ. of Kaiserslautern (Germany)

SESSION 8 MATH RECOGNITION

- 8658 0Z **Structural analysis of online handwritten mathematical symbols based on support vector machines** [8658-23]
F. Simistira, ILSP / "Athena" RIC (Greece) and National Technical Univ. of Athens (Greece); V. Papavassiliou, V. Katsouros, ILSP / "Athena" RIC (Greece); G. Carayannis, National Technical Univ. of Athens (Greece)
- 8658 10 **Using online handwriting and audio streams for mathematical expressions recognition: a bimodal approach** [8658-24]
S. Medjkoune, IRCCyN Lab., Nantes Univ. (France) and LIUM Lab., Le Mans Univ. (France); H. Mouchère, IRCCyN Lab., Nantes Univ. (France); S. Petitrenaud, LIUM Lab., Le Mans Univ. (France); C. Viard-Gaudin, IRCCyN Lab., Nantes Univ. (France)

SESSION 9 INFORMATION RETRIEVAL

- 8658 11 **Using clustering and a modified classification algorithm for automatic text summarization** [8658-26]
A. Aries, H. Oufaida, Ecole Nationale Supérieure d'Informatique (Algeria); O. Nouali, Ctr. de recherche sur l'Information Scientifique et Technique (Algeria)
- 8658 12 **Evaluating supervised topic models in the presence of OCR errors** [8658-27]
D. Walker, E. Ringger, K. Seppi, Brigham Young Univ. (United States)
- 8658 13 **Rule-based versus training-based extraction of index terms from business documents: how to combine the results** [8658-28]
D. Schuster, M. Hanke, K. Muthmann, D. Esser, TU Dresden (Germany)
- 8658 14 **Post processing with first- and second-order hidden Markov models** [8658-29]
K. Taghva, S. Poudel, S. Malreddy, Univ. of Nevada, Las Vegas (United States)
- 8658 15 **Combining discriminative SVM models for the improved recognition of investigator names in medical articles** [8658-30]
X. Zhang, J. Zou, D. X. Le, G. R. Thoma, National Library of Medicine (United States)

SESSION 10 EVALUATION

- 8658 16 **Adaptive detection of missed text areas in OCR outputs: application to the automatic assessment of OCR quality in mass digitization projects** [8658-31]
A. Ben Salah, Bibliothèque Nationale de France (France) and LITIS-Univ. of Rouen (France); N. Ragot, LI-Univ. François Rabelais Tours (France); T. Paquet, LITIS-Univ. of Rouen (France)

- 8658 17 **Evaluating structural pattern recognition for handwritten math via primitive label graphs**
[8658-32]
R. Zanibbi, Rochester Institute of Technology (United States); H. Mouchère,
C. Viard-Gaudin, L'UNAM, IRCCyN, Univ. de Nantes (France)
- 8658 18 **WFST-based ground truth alignment for difficult historical documents with text modification
and layout variations** [8658-33]
M. Al Azawi, Technische Univ. Kaiserslautern (Germany); M. Liwicki, German Research Ctr.
for Artificial Intelligence (Germany); T. M. Breuel, Technische Univ. Kaiserslautern (Germany)

Author Index

Conference Committee

Symposium Chair

Gaurav Sharma, University of Rochester (United States)

Symposium Cochair

Sergio R. Goma, Qualcomm Inc. (United States)

Conference Chairs

Richard Zanibbi, Rochester Institute of Technology (United States)

Bertrand Coüasnon, Institut National des Sciences Appliquées de
Rennes (France)

Conference Program Committee

Gady Agam, Illinois Institute of Technology (United States)

Elisa H. Barney Smith, Boise State University (United States)

William A. Barrett, Brigham Young University (United States)

Kathrin Berkner, Ricoh Innovations, Inc. (United States)

Hervé Déjean, Xerox Research Center 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)

Christopher Kermorvant, A2iA SA (France)

Laurence Likforman-Sulem, Telecom ParisTech (France)

Xiaofan Lin, A9.com, Inc. (United States)

Marcus Liwicki, Deutsches Forschungszentrum für Künstliche
Intelligenz GmbH (Germany)

Daniel P. Lopresti, Lehigh University (United States)

Umapada Pal, India Statistical Institute (India)

Hiroshi Sako, Hosei University (Japan)

Sargur N. Srihari, University at Buffalo, SUNY (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)

Berrin Yanikoglu, Sabanci University (Turkey)

Jie Zou, National Library of Medicine (United States)

Additional Paper Reviewers

Robert Clawson
Douglas Kennard
Jin Chen
Ryuji Mine
Toshinori Miyoshi
Takashi Watanabe

Introduction

On behalf of the DRR XX (2013) Program Committee, welcome to the Twentieth Document Recognition and Retrieval conference being held in San Francisco, California, United States. DRR is held annually as part of the IS&T/SPIE Symposium on Electronic Imaging. It is one of the leading international document recognition conferences, with a growing presence for related information retrieval research.

This year marks the 20th anniversary of the conference, which was first Co-chaired by Luc Vincent and Theo Pavlidis in 1994. There will be a celebration on the first evening of the conference, where there will be sushi and refreshments to commemorate this important event. We thank Google for making this possible, and Brian Hirashiki for his assistance.

A record number of 59 paper submissions were made this year. The previous record was set in 2011 (DRR XVIII) with 54 submissions. 41 papers were accepted (69%), of which 31 were selected for oral presentation (53%, down from 59% last year), and 10 for poster presentation (17%, identical to last year). Our sincerest thanks to the Program Committee and additional referees for helping us create a strong technical program.

For the Best Student Paper Award, 12 authors have submitted 13 papers. We thank Berrin Yanikoglu (chair) and the selection committee for carrying out the difficult task of choosing the winning paper. The winner will be announced in the EI Symposium-wide award ceremony on Wednesday morning of the conference. Google has provided \$500 for the Best Student Paper Award again this year, and we are truly grateful for their continued support of the conference.

We are fortunate to have an excellent pair of keynote talks lined up. Ray Smith of Google Research will provide an account of the evolution of Tesseract: "History of the Tesseract OCR Engine: What Worked and What Didn't (How to Build a World-Class OCR Engine in Less than 20 Years)." And Marti Hearst (University of California, Berkeley), an expert in the areas of information retrieval and search user interfaces will give a talk entitled, "What Does the Future Hold for Search User Interfaces?"

We wish to thank the professionals at SPIE for their assistance with coordinating paper submissions, conference scheduling and resources, and for preparing the final proceedings, and for help with archiving DRR in the DBLP online paper database. Thanks as well to Diana Gonzalez for helping prepare for our anniversary celebration.

We hope that you have an enjoyable and engaging time at DRR this year. Here's to another twenty years!

Richard Zanibbi
Bertrand Coüasnon

xi