

PROCEEDINGS

IS&T / SPIE  
**Electronic  
Imaging**  
SCIENCE AND TECHNOLOGY

# ***Document Recognition and Retrieval XXI***

**Bertrand Coüasnon**

**Eric K. Ringger**

*Editors*

**5–6 February 2014**

**San Francisco, California, United States**

*Sponsored by*

IS&T—The Society for Imaging Science and Technology

SPIE

Published by

SPIE

**Volume 9021**

Proceedings of SPIE 0277-786X, v. 9021

Document Recognition and Retrieval XXI, edited by Bertrand Coüasnon, Eric K. Ringger,  
Proc. of SPIE-IS&T Electronic Imaging, Vol. 9021, 902101 · © 2014 SPIE-IS&T  
CCC code: 0277-786X/14/\$18 · doi: 10.1117/12.2056875

SPIE-IS&T/ Vol. 9021 902101-1

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 XXI*, edited by Bertrand Coüasnon, Eric K. Ringger, Proceedings of SPIE-IS&T Electronic Imaging, SPIE Vol. 9021. Article CID Number (2014)

ISSN: 0277-786X

ISBN: 9780819499387

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 © 2014, 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](http://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/14/\$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 *Conference Committee*
- ix *Introduction*

---

## HANDWRITING

---

- 9021 02 **Writer identification on historical Glagolitic documents** [9021-1]  
S. Fiel, F. Hollaus, M. Gau, R. Sablatnig, Technische Univ. Wien (Austria)
- 9021 03 **Probabilistic modeling of children's handwriting** [9021-2]  
M. Puri, S. N. Srihari, Univ. at Buffalo (United States); L. Hanson, Minnesota Bureau of Criminal Apprehension (United States)
- 9021 04 **Variational dynamic background model for keyword spotting in handwritten documents** [9021-3]  
G. Kumar, S. Wshah, V. Govindaraju, Univ. at Buffalo (United States)
- 9021 05 **Boosting bonsai trees for handwritten/printed text discrimination** [9021-4]  
Y. Ricquebourg, C. Raymond, B. Poirriez, Univ. Européenne de Bretagne, IRISA / INSA de Rennes (France); A. Lemaitre, IRISA, Univ. de Rennes-2 (France); B. Coüasnon, Institut National des Sciences Appliquées de Rennes (France)

---

## FORM CLASSIFICATION

---

- 9021 06 **Form similarity via Levenshtein distance between ortho-filtered logarithmic ruling-gap ratios** [9021-5]  
G. Nagy, Rensselaer Polytechnic Institute (United States); D. Lopresti, Lehigh Univ. (United States)
- 9021 07 **Form classification and retrieval using bag of words with shape features of line structures** [9021-6]  
F. Kleber, M. Diem, R. Sablatnig, Technische Univ. Wien (Austria)

---

## TEXT RECOGNITION

---

- 9021 09 **Utilizing web data in identification and correction of OCR errors** [9021-8]  
K. Taghva, S. Agarwal, Univ. of Nevada, Las Vegas (United States)
- 9021 0A **How well does multiple OCR error correction generalize?** [9021-9]  
W. B. Lund, E. K. Ringger, Brigham Young Univ. (United States); D. D. Walker, Microsoft Corp. (United States)

- 9021 OB **Video text localization using wavelet and shearlet transforms** [9021-10]  
P. Banerjee, Society for Natural Language Technology Research (India); B. B. Chaudhuri, Indian Statistical Institute (India)

---

#### HANDWRITTEN TEXT LINE SEGMENTATION

---

- 9021 OC **A Markov chain based line segmentation framework for handwritten character recognition** [9021-11]  
Y. Wu, Raytheon BBN Technologies (United States); S. Zha, Raytheon BBN Technologies (United States) and Northwestern Univ. (United States); H. Cao, D. Liu, Raytheon BBN Technologies (United States); P. Natarajan, The Univ. of Southern California (United States)
- 9021 OD **Handwritten text segmentation using blurred image** [9021-12]  
A. Lemaitre, IRISA, Univ. Rennes 2 (France); J. Camillerapp, B. Coüason, Institut National des Sciences Appliquées de Rennes (France)

---

#### LAYOUT ANALYSIS

---

- 9021 OF **Optical music recognition on the International Music Score Library Project** [9021-14]  
C. Raphael, R. Jin, Indiana Univ. (United States)
- 9021 OG **Document flow segmentation for business applications** [9021-15]  
H. Daher, LORIA (France); A. Belaid, LORIA, Univ. de Lorraine (France)
- 9021 OH **LearnPos: a new tool for interactive learning positioning** [9021-16]  
C. Carton, Univ. Européenne de Bretagne, IRISA / INSA de Rennes (France); A. Lemaitre, Univ. Européenne de Bretagne, IRISA, Univ. Rennes 2 (France); B. Coüason, Institut National des Sciences Appliquées de Rennes (France)
- 9021 OI **Document page structure learning for fixed-layout e-books using conditional random fields** [9021-17]  
X. Tao, Peking Univ. (China); Z. Tang, C. Xu, Peking Univ. (China) and State Key Lab. of Digital Publishing Technology (China)
- 9021 OJ **Automatic comic page image understanding based on edge segment analysis** [9021-18]  
D. Liu, Y. Wang, Z. Tang, L. Li, L. Gao, Peking Univ. (China)

---

#### INFORMATION RETRIEVAL

---

- 9021 OK **Scalable ranked retrieval using document images** [9021-19]  
R. Jain, D. W. Oard, D. Doermann, Univ. of Maryland, College Park (United States)
- 9021 OL **A contour-based shape descriptor for biomedical image classification and retrieval** [9021-20]  
D. You, S. Antani, D. Demner-Fushman, G. R. Thoma, National Institutes of Health (United States)
- 9021 OM **Semi-automated document image clustering and retrieval** [9021-21]  
M. Diem, F. Kleber, S. Fiel, R. Sablatnig, Technische Univ. Wien (Austria)

- 9021 ON **Fast structural matching for document image retrieval through spatial databases** [9021-22]  
H. Gao, M. Rusiñol, D. Karatzas, J. Lladós, Univ. Autònoma de Barcelona (Spain)

---

#### DATA SETS AND GROUND-TRUTHING

---

- 9021 OO **The Lehigh Steel Collection: a new open dataset for document recognition research** [9021-23]  
B. Bruno, D. Lopresti, Lehigh Univ. (United States)

---

#### INTERACTIVE PAPER SESSION

---

- 9021 OP **Two-stage approach to keyword spotting in handwritten documents** [9021-24]  
M. Haji, IMDS Software (Canada) and Concordia Univ. (Canada); M. R. Ameri, T. D. Bui, C. Y. Suen, Concordia Univ. (Canada); D. Ponson, IMDS Software (Canada)
- 9021 OQ **Extraction and labeling high-resolution images from PDF documents** [9021-25]  
S. K. Chachra, Z. Xue, S. Antani, D. Demner-Fushman, G. R. Thoma, U.S. National Library of Medicine (United States)
- 9021 OR **Structure analysis for plane geometry figures** [9021-26]  
T. Feng, Peking Univ. (China); X. Lu, Peking Univ. (China) and Peking Univ. Founder Group Corp. (China); L. Liu, K. Li, Peking Univ. (China); Z. Tang, Peking Univ. (China) and Peking Univ. Founder Group Corp. (China)
- 9021 OS **On-line signature verification method by Laplacian spectral analysis and dynamic time warping** [9021-27]  
C. Li, L. Peng, C. Liu, X. Ding, Tsinghua Univ. (China)
- 9021 OT **A slant removal technique for document page** [9021-28]  
E. Kavallieratou, Univ. of the Aegean (Greece)
- 9021 OU **Robust binarization of degraded document images using heuristics** [9021-30]  
J. Parker, Georgetown Univ. (United States) and Johns Hopkins Univ. (United States); O. Frieder, G. Frieder, Georgetown Univ. (United States)
- 9021 OV **A machine learning based lecture video segmentation and indexing algorithm** [9021-31]  
D. Ma, B. Xie, G. Agam, Illinois Institute of Technology (United States)

*Author Index*

# Conference Committee

## *Symposium Chair*

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

## *Symposium Cochair*

**Sheila S. Hemami**, Northeastern University (United States)

## *Conference Chairs*

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

**Eric K. Ringger**, Brigham Young University (United States)

## *Conference Program Committee*

**Gady Agam**, Illinois Institute of Technology (United States)

**Sameer K. Antani**, National Library of Medicine (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 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)

**Jiaying Hu**, IBM Thomas J. Watson Research Center (United States)

**Ergina Kavallieratou**, University of the Aegean (Greece)

**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**, Indian Statistical Institute (India)

**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)

**Berrin Yanikoglu**, Sabanci University (Turkey)

**Richard Zanibbi**, Rochester Institute of Technology (United States)

**Jie Zou**, National Library of Medicine (United States)

*Additional Paper Reviewers*

**Alireza Alaei**  
**Sukalpa Chanda**  
**Rajiv Jain**  
**Le Kang**  
**Jayant Kumar**  
**William B. Lund**  
**Varun Manjunatha**  
**Palaiahnakote Shivakumara**

*Session Chairs*

- 1 Handwriting  
**Eric K. Ringger**, Brigham Young University (United States)
  - 2 Form Classification  
**Gady Agam**, Illinois Institute of Technology (United States)
  - 3 Invited Presentation I  
**Bertrand Coüasnon**, Institut National des Sciences Appliquées de Rennes (France)
  - 4 Text Recognition  
**Sameer Antani**, National Library of Medicine (United States)
  - 5 Handwritten Text Line Segmentation  
**Elisa H. Barney Smith**, Boise State University (United States)
  - 6 Invited Presentation II  
**Eric K. Ringger**, Brigham Young University (United States)
  - 7 Layout Analysis  
**Daniel P. Lopresti**, Lehigh University (United States)
  - 8 Information Retrieval  
**Xiaofan Lin**, A9.com, Inc. (United States)
  - 9 Data Sets and Ground-Truthing  
**Bertrand Coüasnon**, Institut National des Sciences Appliquées de Rennes (France)  
**Eric K. Ringger**, Brigham Young University (United States)
- Panel Discussion: Data Sets and Ground-Truthing  
**Bertrand Coüasnon**, *Moderator*, Institut National des Sciences Appliquées de Rennes (France)  
**Eric K. Ringger**, *Moderator*, Brigham Young University (United States)

## Introduction

On behalf of the Document Recognition and Retrieval XXI 2014 (DRR XXI) Program Committee, welcome to the Twenty-first Document Recognition and Retrieval conference being held in San Francisco, California, USA. DRR is held annually as part of the IS&T/SPIE Symposium on Electronic Imaging. It is one of the leading international conferences on document recognition, with a presence for related research on information retrieval and text mining.

This year we received 37 paper submissions. 28 papers were accepted, for an overall acceptance rate of 76%. Of the accepted papers, 21 were selected for oral presentation (57%), and 7 were selected for poster presentation (19%). We want to sincerely thank the Program Committee members and additional referees for helping us create a strong technical program. This year's program includes excellent tracks on Handwriting, Form Classification, Text Recognition, Handwritten Text Line Segmentation, Layout Analysis, Information Retrieval, and Data Sets and Ground-Truthing.

For the Best Student Paper Award, 8 authors have applied. We are grateful to Elisa H. Barney Smith (chair) and the award 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 for the third year, and we are truly grateful for their continued support of the conference.

This year we have two very interesting invited presentations. Ashok Popat and Ray Smith of Google Research will give a joint presentation on "OCR for Google Books" where many challenges arise from the scale and the diverse nature of the scanned corpus. Alexei A. Efros from the University of California, Berkeley, will give a talk entitled, "What makes Big Visual Data Hard?" and speak about problems encountered in collecting and using large visual data sets, based on his extensive research in computer vision.

We hope that you all have an excellent experience at DRR XXI!

**Bertrand Coüasnon**  
**Eric K. Ringger**