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

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

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

HANDWRITING

FORM CLASSIFICATION

TEXT RECOGNITION

HANDWRITTEN TEXT LINE SEGMENTATION

LAYOUT ANALYSIS

INFORMATION RETRIEVAL

DATA SETS AND GROUND-TRUTHING

INTERACTIVE PAPER SESSION

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)
Jianying 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