PROCEEDINGS
IS&T SPIE. ELECTRONIC IMAGING

Document Recognition and Retrieval XXII

Eric K. Ringger
Bart Lamiroy
Editors

11–12 February 2015
San Francisco, California, United States

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

Published by
SPIE

Volume 9402

Document Recognition and Retrieval XXII, edited by Eric K. Ringger, Bart Lamiroy
Proc. of SPIE-IS&T Electronic Imaging, Vol. 9402, 940201 © 2015 SPIE-IS&T
CCC code: 0277-786X/15/$18 · doi: 10.1117/12.2185612

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 XXII*, edited by Eric K. Ringger, Bart Lamiroy, Proceedings of SPIE-IS&T Electronic Imaging, SPIE-IS&T Vol. 9402, Article CID Number (2015)

ISSN: 0277-786X
ISBN: 9781628414929

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. Papers are published as they are submitted and meet publication criteria. A unique 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.

Contents

Authors

Numbers in the index correspond to the last two digits of the six-digit citation identifier (CID) article numbering system used in Proceedings of SPIE. The first four digits reflect the volume number. Base 36 numbering is employed for the last two digits and indicates the order of articles within the volume. Numbers start with 00, 01, 02, 03, 04, 05, 06, 07, 08, 09, 0A, 0B...0Z, followed by 10-1Z, 20-2Z, etc.

Conference Committee

Symposium Chair

Sheila S. Hemami, Northeastern University (United States)

Symposium Co-chair

Choon-Woo Kim, Inha University (Korea, Republic of)

Conference Chairs

Eric K. Ringger, Brigham Young University (United States)
Bart Lamiroy, Université de Lorraine (France)

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)
Bertrand Coüasnon, Institut National des Sciences Appliquées de Rennes (France)
Hervé Déjean, Xerox Research Centre Europe Grenoble (France)
Xiaoqing Ding, Tsinghua University (China)
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, Télécom 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 Research - India (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)
Pingping Xiu, Microsoft Corporation (United States)
Berrin Yanikoglu, Sabanci University (Turkey)
Richard Zanibbi, Rochester Institute of Technology (United States)
Jie Zou, National Library of Medicine (United States)

Conference Review Committee

Jin Chen, Lehigh University (United States)

Session Chairs

Keynote Session I
Bart Lamiroy, Université de Lorraine (France)
Eric K. Ringger, Brigham Young University (United States)

Keynote Session II
Bart Lamiroy, Université de Lorraine (France)
Eric K. Ringger, Brigham Young University (United States)

1 Document Layout Analysis and Understanding
Richard Zanibbi, Rochester Institute of Technology (United States)

2 Document Structure Semantics, Forms, and Tables
Xiaofan Lin, A9.com, Inc. (United States)

3 Text Analysis
Eric K. Ringger, Brigham Young University (United States)

4 Handwriting I
Daniel P. Lopresti, Lehigh University (United States)

5 Quality and Compression
William A. Barrett, Brigham Young University (United States)

6 Graphics and Structure
Bart Lamiroy, Université de Lorraine (France)

7 Handwriting II
Daniel P. Lopresti, Lehigh University (United States)

Introduction

On behalf of the program committee, we welcome you to the twenty-second Document Recognition and Retrieval conference (DRR 2015) 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, including related research on information retrieval and text mining, especially as they pertain to document images.

This year we received 44 paper submissions. Of those, 23 papers were accepted, for an overall acceptance rate of 52%. In this year's edition of the conference, all papers will be presented as oral presentations. We want to sincerely thank the program committee members and additional reviewers for helping us create a strong technical program. The additional reviewers: William Lund (BYU, USA), Douglas Kennard (BYU, USA), Oliver Nina (University of Central Florida, USA), Cérès Carton (IRISA, France), Aurélie Lemaitre (IRISA, France), Yann Ricquebourg (IRISA, Rennes, France), Santosh K.C. (NIH-NLM, USA), Jin Chen (Lehigh University, USA).

The program encompasses the following sessions: Document Layout Analysis and Understanding; Document Structure Semantics, Forms, and Tables; Text Analysis; Handwriting I & II; Quality and Compression; and Graphics and Structure. Topics that are current elsewhere in the pattern recognition and machine learning communities, including probabilistic graphical models and deep neural networks, also make important appearances in the technical program.

This year's invited Keynotes will be given by Brewster Kahle, founder of the Internet Archive, and by Dan Klein, professor of computer science at the University of California, Berkeley. Brewster will address the challenges of large document repositories such as the Internet Archive, while Dan's talk will address unsupervised transcription for documents and music. Our invited speakers are helping to build bridges between DRR and neighboring disciplines such as computer vision and NLP.

Once again we recognize the strong contributions of student primary authors. For the best student paper award, we are grateful to Elisa H. Barney Smith (chair) and the award committee for carrying out the difficult task of choosing winning papers. The winners will be announced at the award ceremony of the conference. This year, thanks to the generous sponsorship by A9 and Google, we are able to complement these awards with a generous gift. We are very grateful for this continued support by our sponsors.

This year promises to hold change for DRR. After many years, SPIE and IS&T have announced that they are parting ways in their sponsorship of the annual Electronic

Imaging conference, DRR's umbrella venue. Therefore DRR will feature a panel discussion in which we address the options for future editions.

We hope that attendees have an excellent experience at DRR XXII and that readers of the proceedings find work that helps them to advance the state of the art in document recognition and retrieval!

Eric K. Ringger
Bart Lamiroy