

# ***Document Recognition and Retrieval XIV***

**Xiaofan Lin**  
**Berrin A. Yanikoglu**  
*Chairs/Editors*

**30 January–1 February 2007**  
**San Jose, California, USA**

*Sponsored by*  
IS&T—The Society for Imaging Science and Technology  
SPIE—The International Society for Optical Engineering

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 XIV*, edited by Xiaofan Lin, Berrin A. Yanikoglu, Proceedings of SPIE-IS&T Electronic Imaging, SPIE Vol. 6500, Article CID Number (2007).

ISSN 0277-786X  
ISBN 9780819466136

Copublished by

**SPIE—The International Society for Optical Engineering**

P.O. Box 10, Bellingham, Washington 98227-0010 USA  
Telephone 1 360/676-3290 (Pacific Time) · Fax 1 360/647-1445  
<http://www.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  
<http://www.imaging.org>

Copyright © 2007, The 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 SPIE and IS&T subject to payment of copying fees. The Transactional Reporting Service base fee for this volume is \$15.00 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 <http://www.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/07/\$15.00.

Printed in the United States of America.

# Contents

vii	<i>Conference Committee</i>
ix	<i>Introduction</i>

---

## SESSION 1 INVITED PAPER I

---

650002	<b>Industrial OCR approaches: architecture, algorithms, and adaptation techniques (Invited Paper)</b> [6500-01] I. Marosi, Nuance-Recognita Corp. (Hungary)
--------	--

---

## SESSION 2 CLASSIFIERS

---

650003	<b>Scale-controlled area difference shape descriptor</b> [6500-02] M. Yang, K. Kpalma, J. Ronsin, IETR, CNRS, INSA de Rennes (France)
650004	<b>Frequency coding: an effective method for combining dichotomizers</b> [6500-03] S. Andra, G. Nagy, Rensselaer Polytechnic Institute (USA); C.-L. Liu, Institute of Automation (China)
650005	<b>A multi-evidence multi-engine OCR system</b> [6500-04] I. Zavorin, E. Borovikov, A. Borovikov, CACI International Inc. (USA); L. Hernandez, Army Research Lab. (USA); K. Summers, M. Turner, CACI International Inc. (USA)

---

## SESSION 3 OCR

---

650006	<b>Interaction for style-constrained OCR</b> [6500-05] S. Veeramachaneni, ITC-IRST (Italy); G. Nagy, Rensselaer Polytechnic Institute (USA)
650007	<b>Reading text in consumer digital photographs</b> [6500-06] V. Vanhoucke, S. B. Gokturk, Riya (USA)

---

**Pagination:** 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. 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.

- 650008 **Adding contextual information to improve character recognition on the Archimedes Palimpsest** [6500-07]  
D. J. Walvoord, R. L. Easton, Jr., R. L. Canosa, Rochester Institute of Technology (USA)
- 650009 **OCR result optimization based on pattern matching** [6500-08]  
J. Shang, C. Liu, X. Ding, Tsinghua Univ. (China)

---

#### SESSION 4 IMAGE PROCESSING

---

- 65000A **Shape from parallel geodesics for distortion correction of digital camera document images** [6500-09]  
K. Fujimoto, Fujitsu Labs. Ltd. (Japan); J. Sun, Fujitsu R&D Ctr. Co., Ltd. (China); H. Takebe, M. Suwa, S. Naoi, Fujitsu Labs. Ltd. (Japan)
- 65000B **Multispectral pattern recognition applied to x-ray fluorescence images of the Archimedes Palimpsest** [6500-10]  
D. M. Hansen, R. L. Easton, Jr., R. Raqueño, Rochester Institute of Technology (USA)
- 65000C **Degraded document image enhancement** [6500-11]  
G. Agam, G. Bal, Illinois Institute of Technology (USA); G. Frieder, The George Washington Univ. (USA); O. Frieder, Illinois Institute of Technology (USA)

---

#### SESSION 5 HANDWRITING RECOGNITION

---

- 65000D **Curvelets based feature extraction of handwritten shapes for ancient manuscripts classification** [6500-12]  
G. Joutel, V. Eglin, S. Bres, H. Emptoz, LIRIS, INSA de Lyon (France)
- 65000E **Interactive training for handwriting recognition in historical document collections** [6500-13]  
D. J. Kennard, W. A. Barrett, Brigham Young Univ. (USA)
- 65000F **Online handwritten mathematical expression recognition** [6500-14]  
H. Büyükbayrak, B. Yanikoglu, A. Erçil, Sabanci Univ. (Turkey)
- 65000G **Recognition of degraded handwritten digits using dynamic Bayesian networks** [6500-15]  
L. Likforman-Sulem, M. Sigelle, Ecole Nationale Supérieure des Télécommunications, TSI (France) and CNRS LTCI (France)

---

#### SESSION 6 DIGITAL PUBLISHING SPECIAL SESSION I

---

- 65000H **Google Books: making the public domain universally accessible** [6500-16]  
A. Langley, D. S. Bloomberg, Google Inc. (USA)
- 65000I **Pixel and semantic capabilities from an image-object based document representation** [6500-17]  
M. Gormish, K. Berkner, M. Boliek, G. Feng, E. L. Schwartz, Ricoh Innovations, Inc. (USA)
- 65000J **Presentation of structured documents without a style sheet** [6500-18]  
S. J. Harrington, E. Wayman, Xerox Corp. (USA)

65000K **Content selection based on compositional image quality** [6500-19]  
P. Obrador, Hewlett-Packard Labs. (USA)

---

**SESSION 7 DIGITAL PUBLISHING SPECIAL SESSION II**

---

65000L **Generic architecture for professional authoring environments to export XML-based formats** [6500-20]  
F. Giannetti, Hewlett-Packard Labs. (United Kingdom)

65000M **Cost-estimating for commercial digital printing** [6500-21]  
M. G. Keif, California Polytechnic State Univ. (USA)

---

**SESSION 8 INFORMATION EXTRACTION AND RETRIEVAL I**

---

65000N **A novel approach for nonuniform list fusion** [6500-23]  
W.-Q. Yan, Univ. of California, Irvine (USA)

65000O **Identification of comment-on sentences in online biomedical documents using support vector machines** [6500-24]  
I. C. Kim, D. X. Le, G. R. Thoma, National Library of Medicine (USA)

65000P **Combining text clustering and retrieval for corpus adaptation** [6500-25]  
F. He, X. Ding, Tsinghua Univ. (China)

---

**SESSION 9 INVITED PAPER II**

---

65000Q **Document recognition serving people with disabilities (Invited Paper)** [6500-26]  
J. R. Fruchterman, The Benetech Initiative (USA)

---

**SESSION 10 INFORMATION EXTRACTION AND RETRIEVAL II**

---

65000R **Title extraction and generation from OCR'd documents** [6500-27]  
K. Taghva, A. Condit, S. Lumos, J. Borsack, T. Nartker, Univ. of Nevada, Las Vegas (USA)

65000S **Content-based document image retrieval in complex document collections** [6500-28]  
G. Agam, S. Argamon, O. Frieder, D. Grossman, Illinois Institute of Technology (USA);  
D. Lewis, David D. Lewis Consulting (USA)

---

**SESSION 11 SEGMENTATION**

---

65000T **A statistical approach to line segmentation in handwritten documents** [6500-29]  
M. Arivazhagan, H. Srinivasan, S. Srihari, Univ. at Buffalo, SUNY (USA)

65000U **Segmentation and labeling of documents using conditional random fields** [6500-30]  
S. Shetty, H. Srinivasan, M. Beal, S. Srihari, Univ. at Buffalo, SUNY (USA)

65000V **Online medical journal article layout analysis** [6500-31]  
J. Zou, D. Le, G. R. Thoma, National Library of Medicine (USA)

65000W **Transcript mapping for handwritten Arabic documents** [6500-32]  
L. M. Lorigo, V. Govindaraju, Univ. at Buffalo (USA)

65000X **Document image content inventories** [6500-33]  
H. S. Baird, M. A. Moll, C. An, M. R. Casey, Lehigh Univ. (USA)

*Author Index*

# Conference Committee

## *Symposium Chairs*

**Michael A. Kriss**, Consultant (USA)  
**Robert A. Sprague**, Consultant (USA)

## *Conference Chairs*

**Xiaofan Lin**, Riya Inc. (USA)  
**Berin A. Yanikoglu**, Sabanci University (Turkey)

## *Program Committee*

**Jan P. Allebach**, Purdue University (USA)  
**Tim L. Andersen**, Boise State University (USA)  
**Apostolos Antonacopoulos**, University of Salford (United Kingdom)  
**Elisa H. Barney Smith**, Boise State University (USA)  
**Kathrin Berkner**, Ricoh Innovations, Inc. (USA)  
**Hui Chao**, Hewlett-Packard Company (USA)  
**Brian D. Davison**, Lehigh University (USA)  
**Xiaoqing Ding**, Tsinghua University (China)  
**David S. Doermann**, University of Maryland, College Park (USA)  
**Steven J. Harrington**, Xerox Corporation (USA)  
**Jianying Hu**, IBM Thomas J. Watson Research Center (USA)  
**Matthew F. Hurst**, Intelliseek, Inc. (USA)  
**Hisashi Ikeda**, Hitachi, Ltd. (Japan)  
**Tapas Kanungo**, IBM Almaden Research Center (USA)  
**Daniel P. Lopresti**, Lehigh University (USA)  
**Thomas A. Nartker**, University of Nevada, Las Vegas (USA)  
**Sargur N. Srihari**, SUNY, University at Buffalo (USA)  
**Kazem Taghva**, University of Nevada, Las Vegas (USA)  
**George R. Thoma**, National Library of Medicine (USA)



## Introduction

Information manifests itself in various media and many forms. The goal of the Document Recognition and Retrieval Conference is to provide a forum to bring together researchers working on all stages and various aspects of converting paper documents into searchable and repurposable media, spanning a wide-area including document recognition (converting document images into electronic form), keyword spotting (retrieving document images containing keywords), information extraction and retrieval from documents, and document repurposing.

The papers in this volume represent the state-of-the-art in diverse areas such as handwriting recognition, optical character recognition, document image processing, information retrieval, and classification. In addition, this year we have six papers in the Digital Publishing Special Sessions.

We are fortunate to have two invited talks by eminent veterans in OCR industry. Dr. István Marosi of Nuance Communications will talk on "Industrial OCR approaches: architecture, algorithms, and adaptation techniques" and Mr. Jim Fruchterman of Benetech will talk on "Document recognition serving people with disabilities."

New for this year, the Best Student Paper will be selected among papers whose lead authors are full-time students. We gratefully acknowledge Nuance Communications, the worldwide leading provider of imaging and speech technologies and solutions, for generously sponsoring this award.

We thank the program committee members and the SPIE conference organizers for their help in organizing the conference and for their help in the review process. We especially thank the participating researchers for their contributions to this cross-disciplinary conference.

As always, we want to hear from readers of this volume, and we encourage you to approach the conference co-chairs or any member of our distinguished program committee with suggestions for future improvements.

We hope everyone finds this year's conference stimulating and rewarding, with its unique opportunities for interaction.

**Xiaofan Lin**  
**Berrin A. Yanikoglu**

