Front Matter: Volume 6815
Document Recognition and Retrieval XV

Berrin A. Yanikoglu
Kathrin Berkner
Editors

29–31 January 2008
San Jose, California, USA

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

Cosponsored by
Ricoh Innovations, Inc. (USA)
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:


ISSN 0277-786X
ISBN 9780819469878

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 © 2008, 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/08/$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 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

<table>
<thead>
<tr>
<th>Session 1</th>
<th>Invited Presentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>6815 02</td>
<td>DRR is a teenager (Invited Paper) [6815-01]</td>
</tr>
<tr>
<td></td>
<td>G. Nagy, Rensselaer Polytechnic Institute (USA)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Session 2</th>
<th>Classification and Recognition I</th>
</tr>
</thead>
<tbody>
<tr>
<td>6815 03</td>
<td>Recognition of Arabic handwritten words using contextual character models [6815-02]</td>
</tr>
<tr>
<td></td>
<td>R. El-Hajj, Univ. of Balamand (Lebanon) and GET-Ecole Nationale Supérieure des Télécommunications (France); C. Mokbel, Univ. of Balamand (Lebanon); L. Likforman-Sulem, GET-Ecole Nationale Supérieure des Télécommunications (France)</td>
</tr>
<tr>
<td>6815 04</td>
<td>Combining different classification approaches to improve off-line Arabic handwritten word recognition [6815-03]</td>
</tr>
<tr>
<td></td>
<td>I. Zavorin, E. Borovikov, E. Davis, A. Borovikov, K. Summers, CACI International Inc. (USA)</td>
</tr>
<tr>
<td>6815 05</td>
<td>Writer adaptation in off-line Arabic handwriting recognition [6815-04]</td>
</tr>
<tr>
<td></td>
<td>G. R. Ball, S. N. Srihari, Univ. at Buffalo (USA)</td>
</tr>
<tr>
<td>6815 06</td>
<td>Whole-book recognition using mutual-entropy-driven model adaptation [6815-05]</td>
</tr>
<tr>
<td></td>
<td>P. Xiu, H. S. Baird, Lehigh Univ. (USA)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Session 3</th>
<th>Image Processing and Enhancement</th>
</tr>
</thead>
<tbody>
<tr>
<td>6815 07</td>
<td>Interactive evolutionary computing for the binarization of degenerated handwritten images [6815-06]</td>
</tr>
<tr>
<td></td>
<td>T. van der Zant, L. Schomaker, A. Brink, Univ. of Groningen (Netherlands)</td>
</tr>
<tr>
<td>6815 08</td>
<td>Correlating degradation models and image quality metrics [6815-27]</td>
</tr>
<tr>
<td></td>
<td>D. K. Reed, E. H. Barney Smith, Boise State Univ. (USA)</td>
</tr>
<tr>
<td>6815 09</td>
<td>Ensemble LUT classification for degraded document enhancement [6815-08]</td>
</tr>
<tr>
<td></td>
<td>T. Obafemi-Ajayi, G. Agam, O. Fréder, Illinois Institute of Technology (USA)</td>
</tr>
<tr>
<td>6815 0A</td>
<td>Automatic removal of crossed-out handwritten text and the effect on writer verification and identification [6815-10]</td>
</tr>
<tr>
<td></td>
<td>A. Brink, H. van der Klauw, L. Schomaker, Univ. of Groningen (Netherlands)</td>
</tr>
</tbody>
</table>
A mixed approach to book splitting [6815-11]
L. Gao, Z. Tang, Peking Univ. (China)

Robust line segmentation for handwritten documents [6815-12]
K. Kuzhinjedathu, H. Srinivasan, S. Srihari, Univ. at Buffalo (USA)

Line-touching character recognition based on dynamic reference feature synthesis [6815-13]
Y. Hotta, K. Fujimoto, Fujitsu Labs., Ltd. (Japan)

Word segmentation of off-line handwritten documents [6815-14]
C. Huang, S. N. Srihari, Univ. at Buffalo (USA)

The OCRopus open source OCR system (Invited Paper) [6815-15]
T. M. Breuel, Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (Germany)
and Univ. Kaiserslautern (Germany)

Measuring the impact of character recognition errors on downstream text analysis [6815-16]
D. Lopresti, Lehigh Univ. (USA)

Online writer identification using character prototypes distributions [6815-17]
S. K. Chan, Univ. Tunku Abdul Rahman (Malaysia); C. Viard-Gaudin, Univ. de Nantes (France);
Y. H. Tay, Univ. Tunku Abdul Rahman (Malaysia)

Stroke frequency descriptors for handwriting-based writer identification [6815-18]
B. Dolega, G. Agam, S. Argamon, Illinois Institute of Technology (USA)

Address block localization based on graph theory [6815-19]
D. Gaceb, V. Eglin, F. Lebourgeois, H. Emptoz, Institut National des Sciences Appliquées
de Lyon (France)

Versatile page numbering analysis [6815-20]
H. Déjean, J.-L. Meunier, Xerox Research Ctr. Europe (France)

Segmentation-based retrieval of document images from diverse collections [6815-21]
M. A. Moll, H. S. Baird, Lehigh Univ. (USA)

Transcript mapping for handwritten English documents [6815-22]
D. Jose, A. Bharadwaj, V. Govindaraju, Univ. at Buffalo (USA)
SESSION 8 INFORMATION EXTRACTION AND DOCUMENT RETRIEVAL

6815 0N Word mining in a sparsely labeled handwritten collection [6815-23]
L. R. B. Schomaker, Univ. of Groningen (Netherlands)

6815 0O An OCR based approach for word spotting in Devanagari documents [6815-24]
A. Bhardwaj, S. Kompalli, S. Setlur, V. Govindaraju, Univ. at Buffalo (USA)

6815 0P Extracting a sparsely located named entity from online HTML medical articles using support
vector machine [6815-25]
J. Zou, D. Le, G. R. Thoma, National Library of Medicine (USA)

6815 0Q Exploring use of images in clinical articles for decision support in evidence-based
medicine [6815-26]
S. Antani, D. Demner-Fushman, National Library of Medicine (USA); J. Li, Univ. at Buffalo
(USA); B. V. Srinivasan, Univ. of Maryland, College Park (USA); G. R. Thoma, National Library
of Medicine (USA)

INTERACTIVE PAPER AND SYMPOSIUM DEMONSTRATION SESSION

6815 0R Model-based document categorization employing semantic pattern analysis and local
structure clustering [6815-37]
K. Fume, Toshiba Corp. (Japan); Y. Ishitani, Toshiba Solutions Corp. (Japan)

6815 0S Large scale parallel document image processing [6815-28]
T. van der Zant, L. Schomaker, E. Valentijn, Univ. of Groningen (Netherlands)

6815 0T A mixed approach to auto-detection of page body [6815-29]
L. Gao, Z. Tang, R. Qiu, Peking Univ. (China)

6815 0U Extracting curved text lines using the chain composition and the expanded grouping
method [6815-30]
N. Noi Bai, K. Nam, Chungbuk National Univ. (South Korea); Y. Song, Chungbuk BIT
Research-Oriented Univ. Consortium (South Korea)

6815 0V Achieving high recognition reliability using decision trees and AdaBoost [6815-31]
J. Xiang, X. Tu, Y. Lu, East China Normal Univ. (China); P. S. P. Wang, Northeastern Univ.
(USA)

6815 0W A generic method for structure recognition of handwritten mail documents [6815-32]
A. Lemaitre, J. Camillerapp, B. Coüasnon, IRISA/INSA (France)

6815 0X Hybrid approach combining contextual and statistical information for identifying MEDLINE
citation terms [6815-33]
I. C. Kim, D. X. Le, G. R. Thoma, National Library of Medicine (USA)

6815 0Y Form classification [6815-34]
K. V. U. Reddy, V. Govindaraju, Univ. at Buffalo (USA)
Interactive degraded document enhancement and ground truth generation [6815-35]
G. Bal, G. Agam, O. Frieder, Illinois Institute of Technology (USA); G. Frieder, George Washington Univ. (USA)

Efficient implementation of local adaptive thresholding techniques using integral images [6815-36]
F. Shafait, D. Keysers, German Research Ctr. for Artificial Intelligence (Germany);
T. M. Breuel, Technical Univ. of Kaiserslautern (Germany)

Author Index
Conference Committee

Symposium Chair

Nitin Sampat, Rochester Institute of Technology (USA)

Conference Chairs

Berrin A. Yanikoglu, Sabanci University (Turkey)
Kathrin Berkner, Ricoh Innovations, Inc. (USA)

Program Committee

Tim L. Andersen, Boise State University (USA)
Apostolos Antonacopoulos, University of Salford (United Kingdom)
Elisa H. Barney-Smith, Boise State University (USA)
Xiaoqing Ding, Tsinghua University (China)
David S. Doermann, University of Maryland, College Park (USA)
Jianying Hu, IBM Thomas J. Watson Research Center (USA)
Matthew F. Hurst, Intelliseek, Inc. (USA)
Hisashi Ikeda, Hitachi, Ltd. (Japan)
Tapas Kanungo, Yahoo! Inc. (USA)
Laurence Likforman-Sulem, GET-École Nationale Supérieure des Télécommunications (France)
Xiaofan Lin, Riya Inc. (USA)
Daniel P. Lopresti, Lehigh University (USA)
Lambert R. B. Schomaker, University of Groningen (Netherlands)
Sargur N. Sridhar, University at Buffalo (USA)
Venkata Subramaniam, IBM India Research Laboratory (India)
Kazem Taghva, University of Nevada, Las Vegas (USA)
George R. Thoma, National Library of Medicine (USA)