Lecture Notes in Computer Science 4768

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

David Hutchison
Lancaster University, UK

Takeo Kanade
Carnegie Mellon University, Pittsburgh, PA, USA

Josef Kittler
University of Surrey, Guildford, UK

Jon M. Kleinberg
Cornell University, Ithaca, NY, USA

Alfred Kobsa
University of California, Irvine, CA, USA

Friedemann Mattern
ETH Zurich, Switzerland

John C. Mitchell
Stanford University, CA, USA

Moni Naor
Weizmann Institute of Science, Rehovot, Israel

Oscar Nierstrasz
University of Bern, Switzerland

C. Pandu Rangan
Indian Institute of Technology, Madras, India

Bernhard Steffen
University of Dortmund, Germany

Madhu Sudan
Massachusetts Institute of Technology, MA, USA

Demetri Terzopoulos
University of California, Los Angeles, CA, USA

Doug Tygar
University of California, Berkeley, CA, USA

Gerhard Weikum
Max-Planck Institute of Computer Science, Saarbruecken, Germany

David Doermann Stefan Jaeger (Eds.)

Arabic and Chinese Handwriting Recognition

SACH 2006 Summit
College Park, MD, USA, September 27-28, 2006
Selected Papers

Springer

Volume Editors

David Doermann
Stefan Jaeger
University of Maryland
Institute for Advanced Computer Studies
A.V. Williams Building, College Park, MD 20742, USA
E-mail: {doermann, jaeger}@umd.edu

Library of Congress Control Number: 2008923755

CR Subject Classification (1998): I.5, H.2.4

LNCS Sublibrary: SL 6 – Image Processing, Computer Vision, Pattern Recognition, and Graphics

ISSN 0302-9743
ISBN-10 3-540-78198-6 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-78198-1 Springer Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

Springer is a part of Springer Science+Business Media

springer.com

© Springer-Verlag Berlin Heidelberg 2008
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12228713 06/3180 5 4 3 2 1 0

Preface

In the fall of 2006, the University of Maryland, along with various government and industrial sponsors, invited leading researchers from all over the world to a two-day Summit on Arabic and Chinese Handwriting Recognition (SACH 2006). The event acted as a complement to the biennial Symposium on Document Image Understanding Technology (SDIUT), providing a focused glimpse into the state of the art in Arabic and Chinese handwriting recognition. It offered a forum for interaction with prominent researchers at the forefront of the scientific community and provided an opportunity for participants to help explore possible directions of the field. This book is a result of the expansion, peer review, and revision of selected papers presented at this meeting.

Handwriting recognition remains the Holy Grail of document analysis, and Arabic and Chinese scripts embrace many of the most significant challenges. We are pleased to have 16 scientific papers covering the original topics of handwritten Arabic and Chinese, as well as 2 papers covering other handwritten scripts. We asked each author to not only describe the techniques used in addressing the problem, but to attempt to identify the key research challenges and problems that the community faces. The result is an impressive collection of manuscripts that provide various detailed views of the state of research.

In this book, six articles deal directly with Arabic handwriting.

- Cheriet provides an overview of the problems of Arabic recognition and how systems can use natural language processing techniques to correct errors in lexicon-based systems.
- Suen et al. focus on Persian script, providing a historical survey of the language and techniques used in recognition.
- Belaid et al. provide a hybrid approach that attempts to take advantage of both local and global properties of the language.
- Srihari et al. focus on the problem of search, providing an image-based technique for matching words in Arabic documents.
- Abdulkader introduces language models taking advantage of Part of Arabic Words (PAWS) to aid segmentation.

With the increased interest in Arabic recognition, evaluation is of increased interest as well.

- Märgner and El Abed provide a survey of datasets and competitions that have emerged in recent years for Arabic handwriting recognition.

The history of Chinese handwriting recognition research is much longer than the research record for Arabic handwriting recognition.

- Liu provides an extensive survey of the effects of normalization and feature extraction on recognition of Chinese.
- Fujisawa explores techniques for dealing with uncertainty in the Chinese postal automation domain.
- Guo describes an approach to clustering in a coarse-to-fine hierarchical classification system.
- Chang describes techniques for dealing with the large-scale classification problems of Chinese character recognition.
- Nakagawa et al. discuss the challenges and techniques for the online Chinese character recognition problem.

To contrast techniques developed for Chinese and Arabic, several papers have been included that address cross cutting methods applied to other languages.

- Ding et al. apply a segmentation-driven approach to the recognition of both Chinese and Arabic.
- Lopresti et al. focus on word recognition using language models, introducing two new classifiers that are able to adapt for style-specific applications.
- Natarajan describes a complete segmentation-free system that uses classical Hidden Markov Models.

And finally

- Bunke et al. describe a line-level recognition system applied to handwritten English text, and
- Pal et al. describe their work on the recognition of South Indian handwritten scripts.

With this book, we tried our best to provide you with a meaningful overview of the state of the art and research trends. We sincerely hope that the challenges set forth by these authors motivate other researchers to continue to address these difficult problems.

November 2007

Stefan Jaeger
David Doermann

Table of Contents

Visual Recognition of Arabic Handwriting: Challenges and New Directions 1
Mohamed Cheriet

A Review on Persian Script and Recognition Techniques 22
Sara Izadi, Javad Sadri, Farshid Solimanpour, and Ching Y. Suen

Human Reading Based Strategies for Off-Line Arabic Word Recognition 36
Abdel Belaïd and Christophe Choisy

Versatile Search of Scanned Arabic Handwriting 57
Sargur N. Srihari, Gregory R. Ball, and Harish Srinivasan

A Two-Tier Arabic Offline Handwriting Recognition Based on Conditional Joining Rules 70
Ahmad AbdulKader

Databases and Competitions: Strategies to Improve Arabic Recognition Systems 82
Volker Märgner and Haikal El Abed

Handwritten Chinese Character Recognition: Effects of Shape Normalization and Feature Extraction 104
Cheng-Lin Liu

How to Deal with Uncertainty and Variability: Experience and Solutions 129
Hiromichi Fujisawa

An Efficient Candidate Set Size Reduction Method for Coarse-Classification in Chinese Handwriting Recognition 152
Feng-Jun Guo, Li-Xin Zhen, Yong Ge, and Yun Zhang

Techniques for Solving the Large-Scale Classification Problem in Chinese Handwriting Recognition 161
Fu Chang

Recent Results of Online Japanese Handwriting Recognition and Its Applications 170
Masaki Nakagawa, Junko Tokuno, Bilan Zhu, Motoki Onuma, Hideto Oda, and Akihito Kitadai

Segmentation-Driven Offline Handwritten Chinese and Arabic Script Recognition 196
Xiaoqing Ding and Hailong Liu

Multi-character Field Recognition for Arabic and Chinese Handwriting 218
Daniel Lopresti, George Nagy, Sharad Seth, and Xiaoli Zhang

Multi-lingual Offline Handwriting Recognition Using Hidden Markov Models: A Script-Independent Approach 231
Prem Natarajan, Shirin Saleem, Rohit Prasad, Ehry MacRostie, and Krishna Subramanian

Handwritten Character Recognition of Popular South Indian Scripts 251
Umapada Pal, Nabin Sharma, Tetsushi Wakabayashi, and Fumitaka Kimura

Ensemble Methods to Improve the Performance of an English Handwritten Text Line Recognizer 265
Roman Bertolami and Horst Bunke

Author Index 279