Advances in Pattern Recognition

For further volumes:
http://www.springer.com/series/4205

Venu Govindaraju · Srirangaraj Setlur
Editors

Guide to OCR for Indic Scripts

Document Recognition and Retrieval

Springer

Editors
Prof. Venu Govindaraju
Center for Unified Biometrics and Sensors
520 Lee Entrance
Amherst NY 14228
Suite 202
USA
govind@buffalo.edu
venu.govindaraju@gmail.com

Srirangaraj (Ranga) Setlur
Center for Unified Biometrics and Sensors
520 Lee Entrance
Amherst NY 14228
Suite 202
USA
setlur@buffalo.edu
setlur@gmail.com

Series Editor
Professor Sameer Singh, PhD
Research School of Informatics
Loughborough University
Loughborough, UK

ISBN 978-1-84800-329-3 e-ISBN 978-1-84800-330-9
DOI 10.1007/978-1-84800-330-9
Springer London Dordrecht Heidelberg New York

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Control Number: 2009934526

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Foreword

The original motivations for developing optical character recognition technologies were modest to convert printed text on flat physical media to digital form, producing machine-readable digital content. By doing this, words that had been inert and bound to physical material would be brought into the digital realm and thus gain new and powerful functionalities and analytical possibilities.

First-generation digital OCR researchers in the 1970s quickly realized that by limiting their ambitions primarily to contemporary documents printed in standard font type from the modern Roman alphabet (and of these, mostly English language materials), they were constraining the possibilities for future research and technologies considerably. Domain researchers also saw that the trajectory of OCR technologies if left unchanged would exclude a large portion of the human record. Digital conversion of documents and manuscripts in other alphabets, scripts, and cursive styles was of critical importance. Embedded in non-Roman alphabet source documents, including ancient manuscripts, papyri scrolls, clay tablets, and other inscribed artifacts was not only a wealth of scholarly information but also new opportunities and challenges for advancing OCR, imaging sciences, and other computational research areas. The limiting circumstances at the time included the rudimentary capability (and high cost) of computational resources and lack of network-accessible digital content. Since then computational technology has advanced at a very rapid pace and networking infrastructure has proliferated. Over time, this exponential decrease in the cost of computation, memory, and communications bandwidth combined with the exponential increase in Internet-accessible digital content has transformed education, scholarship, and research. Large numbers of researchers, scholars, and students use and depend upon Internet-based content and computational resources.

The chapters in this book describe a critically important area of investigation – addressing conversion of Indic script into machine-readable form. Rough estimates have it that currently more than a billion people use Indic scripts. Collectively, Indic historic and cultural documents contain a vast richness of human knowledge and experience.

The state-of-the-art research described in this book demonstrates the multiple values associated with these activities. Technically, the problems associated with Indic script recognition are very difficult and will contribute to and inform related

script recognition efforts. The work also has enormous consequence for enriching and enabling the study of Indic cultural heritage materials and the historic record of its people. This in turn broadens the intellectual context for domain scholars focusing on other societies, ancient and modern.

Digital character recognition has brought about another milestone in collective communication by bringing inert, fixed-in-place, text into an interactive digital realm. In doing so, the information has gained additional functionalities which expand our abilities to connect, combine, contextualize, share, and collaboratively pursue knowledge making. High-quality Internet content continues to grow in an explosive fashion. In the new global cyberenvironment, the functionalities and applications of digital information continue to transform knowledge into new understandings of human experience and the world in which we live. The possibilities for the future are limited only by available research resources and capabilities and the imagination and creativity of those who use them.

Arlington, Virginia

Stephen M. Griffin
Program Director
National Science Foundation

Preface

In the past few years, we have seen many ambitious large-scale efforts to make all written material accessible online in a digital format. University-led efforts such as the Million Book Project and industry-led efforts such as the Google Books Library Project have made this goal seem attainable, though there still remains a lot to be achieved. Government sponsored programs such as the Digital Libraries Initiative, which is currently in its second phase, jointly supported by the National Science Foundation (NSF), the Defense Advanced Research Projects Agency (DARPA), the National Library of Medicine (NLM), the Library of Congress (LoC), the National Aeronautics and Space Administration (NASA), the National Endowment for the Humanities (NEH) and others have led to the acceleration of development of technologies and applications that make it possible to create, access and manage digital library content in a quick and easy manner. The Million Book Project was in fact, a product of the Digital Libraries Initiative. Advances in scanning technologies have resulted in the creation of large libraries of digitized content under these initiatives.

The objective of the creation of digital library content is to enable the query and retrieval of relevant documents from the digital library. Technologies such as keyword spotting and optical character recognition (OCR) are crucial elements in the creation of indexed digital library content. The state of the art today is still a long way from being able to reliably recognize the text in many of documents that have been digitized to date. This is especially true in the case of documents in non-Latin scripts and in the case of most historical documents.

Most of the scripts of south and Southeast Asia are derived from the ancient *Brahmi* script (The Unicode Standard v3.0). Since a majority of these scripts are mainly prevalent in the Indian sub-continent, they are also called *Indic* scripts. The basic character set for Indic scripts can be seen in Figs. 1 and 2.

The Eighth Schedule of the Constitution of India contains a list of 22 major languages that are currently used in India. Additionally, there are hundreds of minor languages or dialects that are spoken by populations in small geographical pockets making south Asia a highly multi-lingual region. The scripts used by contemporary speakers of these languages for writing are *Devanagari* (Sanskrit, Hindi, Marathi, Nepali, Konkani, Santhali, Bodo, Dogri, Kashmiri, Maithili, Sindhi), *Bengali* (Bengali or Bangla, Assamese or Asomiya, Manipuri, Santhali), *Gurmukhi*

Fig. 1 Vowels and a consonant with vowel modifiers in Indic scripts

(Punjabi), *Gujarati* (Gujarati), *Oriya* (Oriya, Santhali), *Tamil* (Tamil), *Telugu* (Telugu), *Kannada* (Kannada), *Malayalam* (Malayalam). Many of these languages were also historically written in other related scripts. Urdu, spoken in many parts of India and Pakistan, is usually written in the Perso-Arabic *Nastaliq* script. Some languages such as Kashmiri, Sindhi, Dogri are also written using the Perso-Arabic script in certain regions. Given the widespread use of Urdu in India, we have loosely defined the term *Indic* scripts in the context of this book to include the Perso-Arabic script used for Urdu. Most Indic scripts follow a writing system that is written from left-to-right and has the orthographic syllable as the effective unit consisting of a consonant and vowel core optionally preceded by one or more consonants. The Perso-Arabic script used for Urdu is written from right-to-left.

An enormous body of literature spanning a range of topics that would be of interest to researchers and lay people alike exists in these languages and scripts. There is very little digital access to these contemporary and cultural heritage materials in *Indic* scripts. These scripts present some challenges for OCR which are different from the issues faced with Latin and Oriental scripts. There are also heritage materials in these scripts that are written on media such as palm leaf that pose problems in digitization as well as image pre-processing to render them potentially suitable for down-the-line indexing via keyword spotting or partial or complete OCR.

All major research groups working in the area of *Indic* OCR and information retrieval from *Indic* documents are represented in this book. The groups include

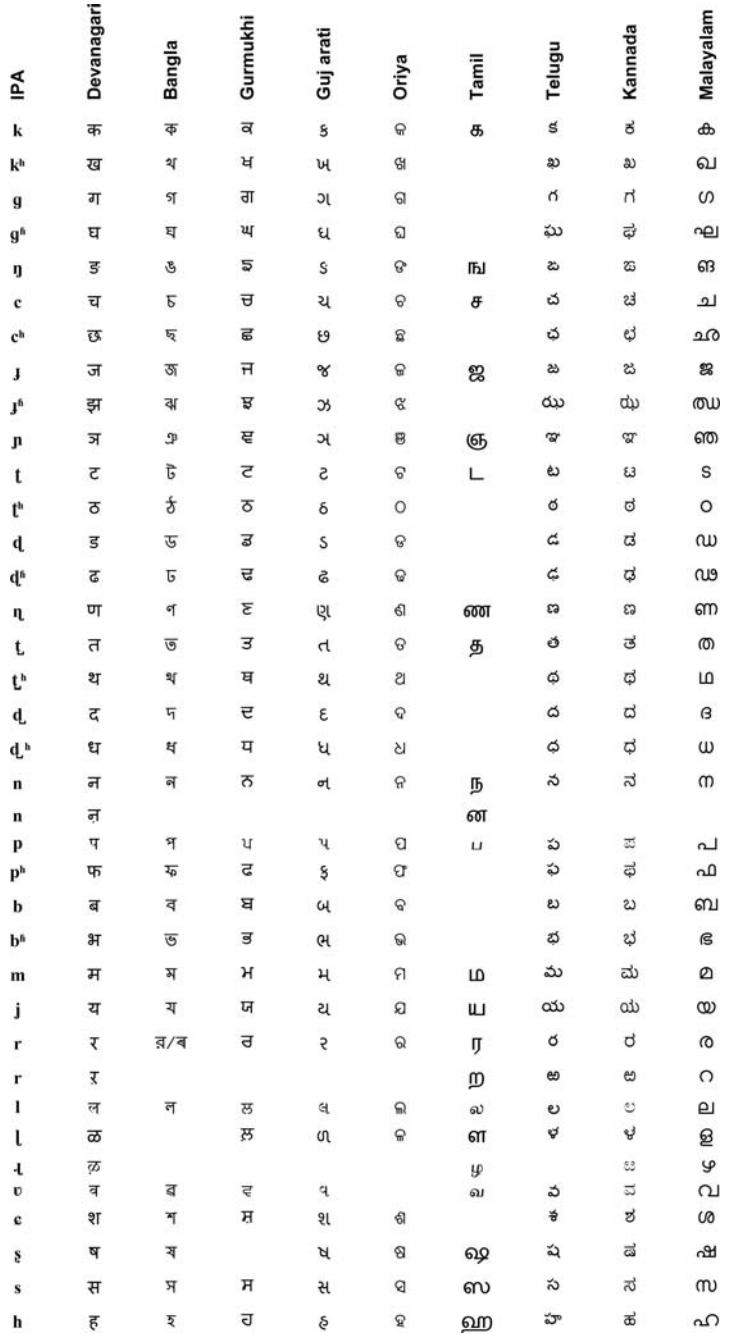

Fig. 2 Consonants in Indic Scripts

university researchers and industry labs in India as well as in the United States. This book is divided into two parts: recognition of Indic scripts and retrieval of Indic documents. The recognition part covers topics such as data set creation for OCR development and the current state of the art in the development of OCR technologies for many of the Indic scripts. We also include a chapter on online handwriting recognition for Indic scripts that is critical for the development of applications built around digital pen input of Indic script. The second part discusses issues related to facilitating query and retrieval of Indic documents from digital libraries.

1 Part I: Recognition of Indic Scripts

Indic scripts are *abugidas* or *alpha-syllabaries*, i.e., the basic writing unit consists of a consonant–vowel core and phonetically, they largely share the same basic character set (vowels and consonants) as shown in Figs. 1 and 2. A vowel has two forms, an independent form when not part of a consonant and a dependent form. In the written form, the manner in which the dependent vowel signs or *maatraas* are attached to the base consonant exhibit a large variation among the Indic scripts. These scripts are also characterized by a large number of consonant conjunct forms where the characters tend to change shape depending on their context. This results in a large set of character glyphs and poses a challenge for OCR systems Indic numerals are shown Fig. 3.

Availability of data sets is a critical requirement for the development of OCR systems. The chapter on data sets describes the ongoing work at IIIT, Hyderabad, on the creation of a large data corpus that currently has over 600,000 document images representing many Indic scripts. This chapter comprehensively details the steps involved in the creation of a good data set including the identification of documents, procedure for scanning and creation of images, consistent procedures

Fig. 3 Numerals in Indic scripts (0 in Tamil and Malayalam are not native)

for annotation, and structured storage of the metadata to allow for efficient indexing and retrieval.

The next nine chapters describe OCR systems that cover eight different scripts: Bangla, Devanagari, Gurmukhi, Gujarati, Kannada, Malayalam, Tamil, and Urdu (Perso-Arabic). It may be noted that while these scripts share some similarities they are also quite disparate. The methods described in these chapters span the use of a multitude of features and classification techniques giving the reader a good insight into the efficacy of these methods for the various Indic scripts.

The work on Bangla and Devanagari OCR at ISI, Kolkata, uses sequential rules to segment characters followed by template matching for classification using a bank of classifiers. The chapter also describes the use of post-processing of recognition results to improve classification performance and a methodology for error evaluation.

A system for recognition of machine-printed Gurmukhi documents has been presented by the researchers at Punjabi University. Local and global structural features are used with a multi-stage classification approach using binary tree and k-nearest neighbor classifiers.

The next chapter describes work on Gujarati documents by researchers at the University of Baroda and the University of Hyderabad. The work explores multiple feature extraction techniques such as fringe maps, discrete cosine transforms and wavelets and multiple classifiers such as a nearest neighbor classifier and a neural network-based classifier. Experimental results are presented comparing various feature–classifier combinations.

The system developed at IISc Bangalore for recognition of bilingual documents (Kannada and English) addresses a frequent challenge encountered in the sub-continent, viz., documents containing multiple scripts. A script identification method based on Gabor filters and discrete cosine transforms is proposed and classification using nearest-neighbor, linear discriminant classifiers, and support vector machines are compared. Graph-based features and an SVM for based classifier have been used for the OCR.

The chapter on Malayalam documents describes work on both machine-printed documents and online handwriting at IIIT Hyderabad. A novel approach has been used to learn features automatically from large quantities of training data, i.e., to derive a statistical feature extraction suitable for the script from examples, rather than defining intuitive features from experience. An ensemble of binary SVM classifiers using decision-directed acyclic graphs is used for classification.

The work on OCR of Tamil magazine documents at IIT Madras includes layout analysis and segmentation of body text, titles, and images using a modified smeared run-length approach. The character recognition is based on a radial basis function neural network and uses Gabor filter features.

The chapter on recognition of Urdu handwriting presents an overview of existing research on Urdu documents and reports preliminary experiments at the University at Buffalo on handwritten Urdu documents using GSC features and k-NN and SVM classifiers.

The BBN Byblos Hindi OCR system uses a script-independent methodology for OCR using hidden markov models. The chapter from BBN Technologies describes

their efforts in training of the system for Hindi (Devanagari) documents and presents experimental results on Hindi documents.

The chapter from University at Maryland describes a novel technique using font models for script identification and segmentation of Hindi characters in machine-printed documents. In the recognition system, three feature extraction methods are used to demonstrate the importance of appropriate features for classification.

The last chapter in this part from HP Labs, Bangalore describes challenges in the recognition of handwriting in Indic scripts in the online domain and provides an overview of the state of the art in isolated character and word recognition. It also describes the progress in the development of applications such as handwriting-based text input systems.

2 Part II: Retrieval of Indic Documents

The first chapter in this part describes ongoing work at Brown University in enhancing access to cultural heritage materials of India for researchers through The Sanskrit Library at Brown. The methodology could be extended to other digital library collections throughout the world.

The next chapter describes techniques developed at the University at Buffalo to enhance images of historical Indic manuscripts such as palm leaf manuscripts and render them at the very least readable for human eyes and potentially enable segmenting of lines of text and even keyword spotting or partial OCR for indexing and retrieval. Novel methods for image enhancement using background normalization and text line location and extraction using an adaptive local connectivity map have been presented.

The following two chapters describe two different techniques for word spotting. The first work of IIT Delhi uses a geometric feature graph to encode word image features for word spotting. The graph is encoded as a string that serves as a compressed representation of the word image skeleton. The GFG-based word image spotting is augmented with latent semantic analysis for more effective retrieval.

The other work on word spotting from the University at Buffalo describes two techniques, a script-dependent, recognition-based approach using a block adjacency graph representation and a script-independent recognition-free approach based on image moments.

The last chapter reviews the state of the art in mono-lingual and cross-lingual information retrieval in Indic languages. A framework for evaluation of Indian language information retrieval has been described.

3 Target Audience

This unique guide/reference is the very first comprehensive book on the subject of OCR for Indic scripts, providing an overview of the state-of-the-art research in this field as well as other issues related to facilitating query and retrieval of Indic

documents from digital libraries. We hope that this guide will serve as an excellent reference for researchers and graduate students studying OCR technology and methodologies for Indic scripts.

Acknowledgments

We are extremely thankful to all the chapter authors who have contributed to making this book an invaluable resource to researchers in the area of OCR for Indic scripts. The 16 chapters are the combined efforts of 37 authors. Interaction with the authors through the concept, draft, and finalization stages has been truly stimulating and enjoyable. We would like to thank Stephen Griffin, Program Director at NSF for his foreword and his continued support for the development of multilingual OCR technologies. We would also like to thank Achint Thomas and Anurag Bhardwaj for helping with submissions, proofreading, and a myriad of other tasks. We are also grateful to Springer for their encouragement and persistence in driving us to complete this work in a timely manner.

Buffalo, New York

Venu Govindaraju
Srirangaraj Setlur

Contents

Contributors

Mudit Agrawal LAMP of UMIACS, University of Maryland, College Park, MD 20742, USA, mudit@umd.edu

A. Bharath Hewlett-Packard Laboratories, Bangalore, India, bharath.a@hp.com

Anurag Bhardwaj Department of Computer Science and Engineering, Center for Unified Biometrics and Sensors, University at Buffalo, Amherst, NY, USA, ab94@cubs.buffalo.edu

Srinivasa Chakravarthy Department of Biotechnology, IIT – Madras, Chennai, India, schakra@ee.iitm.ac.in

B. B. Chaudhuri Indian Statistical Institute, Kolkata, India, bbcisical@gmail.com

Santanu Chaudhury IIT Delhi, New Delhi, India, schaudhury@ee.iitd.ac.in

Michael Decerbo BBN Technologies, Cambridge, MA, USA, mdecerbo@bbn.com

Jignesh Dholakia Department of Applied Mathematics, M S University of Baroda, Vadodara, Gujarat, India, jignesh.dholakia@gmail.com

David Doermann LAMP of UMIACS, University of Maryland, College Park, MD, USA, doermann@umd.edu

Ritu Garg IIT Delhi, New Delhi, India, ritu2721a@gmail.com

Venu Govindaraju Department of Computer Science and Engineering, Center for Unified Biometrics and Sensors, University at Buffalo, Amherst, NY, USA, venu@buffalo.edu

Gaurav Harit IIT Delhi, New Delhi, India, gharit@ee.iitd.ac.in

Malcolm Hyman Department of Classics, Brown University, Providence, RI, USA

C. V. Jawahar International Institute for Information Technology, Center for Visual Information Technology, Hyderabad, India, jawahar@iiit.ac.in

K. J. Jinesh International Institute for Information Technology, Center for Visual Information Technology, Hyderabad, India

Aparna Kokku Department of Biotechnology, IIT – Madras, Chennai, India, ee01m03@ee.iitm.ac.in

Anand Kumar International Institute for Information Technology. Center for Visual Information Technology, Hyderabad, India

Gurpreet Singh Lehal Department of Computer Science, Punjabi University, Patiala, India, gslehal@gmail.com

Huanfeng Ma Drexel Hill, PA 19026, USA, huanfeng@gmail.com

Ehry MacRostie BBN Technologies, Cambridge, MA, USA, emacrost@bbn.com

Sriganesh Madhvanath Hewlett-Packard Laboratories, Bangalore, India, srig@hp.com

Prasenjit Majumder CVPR Unit, Indian Statistical Institute, Kolkata, India, prasenjit_t@isical.ac.in

Mandar Mitra CVPR Unit, Indian Statistical Institute, Kolkata, India, mandar@isical.ac.in

Omar Mukhtar Department of Computer Science and Engineering, Center for Unified Biometrics and Sensors, University at Buffalo, Amherst, NY, USA, omukhtar@buffalo.edu

Anoop Namboodiri International Institute for Information Technology, Center for Visual Information Technology, Hyderabad, India, anoop@iiit.ac.in

P. J. Narayanan International Institute for Information Technology, Center for Visual Information Technology, Hyderabad, India, pjn@iiit.ac.in

Prem Natarajan BBN Technologies, Cambridge, MA, USA, prem@bbn.com

N. V. Neeba International Institute for Information Technology, Center for Visual Information Technology, Hyderabad, India

Atul Negi Department of Computer and Information Sciences, University of Hyderabad, Hyderabad, India, atulcs@uohyd.ernet.in

Peeta Basa Pati Department of Electrical Engineering, Indian Institute of Science, Bangalore, India

A. Phaneendra International Institute for Information Technology, Center for Visual Information Technology, Hyderabad, India

S. Rama Mohan Department of Applied Mathematics, M. S. University of Baroda, Vadodara, Gujarat, India, srmmsu@gmail.com

A. G. Ramakrishnan Department of Electrical Engineering, Indian Institute of Science, Bangalore, India, ramkiag@ee.iisc.ernet.in

Peter M. Scharf Department of Classics, Brown University, Providence, RI, USA, scharf@brown.edu

Srirangaraj Setlur Department of Computer Science and Engineering, Center for Unified Biometrics and Sensors, University at Buffalo, Amherst, NY, USA, setlur@buffalo.edu

Zhixin Shi Department of Computer Science and Engineering, Center for Unified Biometrics and Sensors, University at Buffalo, Amherst, NY, USA, zshi@buffalo.edu

R. S. Umesh Department of Electrical Engineering, Indian Institute of Science, Bangalore, India