SpringerBriefs in Electrical and Computer Engineering

For further volumes:
http://www.springer.com/series/10059

Tonghua Su

# Chinese Handwriting Recognition: An Algorithmic Perspective

Springer

Dr. Tonghua Su
School of Software/School of Computer
Science and Technology
Harbin Institute of Technology
Harbin
Heilongjiang
People's Republic of China

ISSN 2191-8112 ISSN 2191-8120 (electronic)
ISBN 978-3-642-31811-5 ISBN 978-3-642-31812-2 (eBook)
DOI 10.1007/978-3-642-31812-2
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012955476

© The Author(s) 2013
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

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

# Foreword

A study on printed Chinese character recognition was first attempted in the mid-1960s in the United States (Casey and Nagy 1966). In the late-1970s a printed Kanji recognition system was built in Japan where Chinese characters are commonly used as Kanji. Around the same time, the Electrotechnical Laboratory in Japan collected and released a hand printed Kanji database ETL-8 to initiate research on hand printed Chinese character recognition. In the 1980s, significant feature extraction and character recognition techniques such as chain code histogram and MQDF were proposed by a group of Mie University and began to be widely employed in Chinese character recognition. Because of the huge requirements on the storage and processing time the research on Chinese character recognition of those days used to be carried out on mainframe computers or more typically on super computers, while commercial OCR systems were implemented employing hardware technologies such as micro programming, DSP, and dedicated LSI.

Great advances had been witnessed in the constrained handwritten Chinese character recognition and the restricted domain during the 1990s. In the 1990s handwritten address recognition techniques were realized based on the integrated segmentation-recognition strategy for automatic mail sorting in the United States and in the late-1990s mail sorting machines based on the similar strategy for handwritten Kanji address were deployed in Japan. With the rapid progress of personal computers the commercial OCR software for Chinese character recognition was released and became widely available in this period. With increasing industrial demands on further performance improvement and more naturally handwritten samples, extension of the recognition target, e.g., unconstrained handwritten Chinese text, has been raised and activated many researches in this area since then to the present.

This book provides an excellent algorithmic perspective of Chinese handwriting recognition. After a thorough survey on the algorithmic development of Chinese character recognition, the details of the large-scale database, HIT-MW, collected by Harbin Institute of Technology are described. The research on unconstrained handwritten Chinese text can be conducted with the support of HIT-MW database. Then, the integrated segmentation-recognition strategy embedded with MQDF

classifier is constructed incrementally. In particular, an effective and scalable PL-MQDF is devised. Most of all, the deep investigation on HMM-based segmentation-free strategy for Chinese handwriting recognition should be highlighted. The sophiscated algorithms to improve the performance are of significant inspiration to Markov model-based systems. Although segmentation and recognition are often treated conveniently as two different processes, those in human recognition of characters seem to be more concurrent and simultaneous. From my point of view this book is expected to contribute widely to the progress of this research area in the future.

28 October 2012

Prof. Fumitaka Kimura
Graduate School of Engineering
Mie University

## Reference

R. Casey, G. Nagy, Recognition of printed Chinese characters. IEEE Trans. Electron. Comput. EC-**15**(1), 91–101 (1966)

# Preface

Creating a practical Chinese handwriting recognizer from scratch is no mean feat. First of all, you need to ensure the availability of sufficient handwritten samples and/or have to collect your own. It took my group three entire years to compile the HIT-MW database. The design and implementation involves thorough evaluation of the state-of-the-art algorithms, ranging from image preprocessing and feature extraction to classifier training and path decoding. You also need to balance the efficacy and training workload. Thus, this book is devoted to helping interested researchers/engineers/students quickly embark on their study of Chinese handwriting recognition.

This work also reflects my ten years of experience in Chinese handwriting recognition. First, it presents an overview of Chinese handwriting recognition within the context of the character recognition field, with an emphasis on the algorithmic developments and shifts from standard databases. Considering the huge challenges posed by realistic Chinese handwriting recognition, we had to stand on the shoulders of giants. Two strategies, the segmentation-free and integrated segmentation-recognition strategy, are pursued in the literature. Recent developments in these strategies are investigated and algorithms that have worked well in practice are primarily focused on in this book. In order to provide a concrete and valuable reference resource, all algorithms are tested on standard databases.

The book employs a "simple-to-complex" organizing principle. As such, the division of the chapters is not on the basis of system components, e.g., preprocessing, feature extraction, classifier design, etc. Instead, baseline systems are initially presented and are subsequently expanded on and incrementally improved. For better comprehension, the first chapter provides a bird's-eye view and Chap. 2 presents a dedicated scheme for establishing the HIT-MW database. In this way, readers have the freedom to begin with whichever chapter they prefer.

I owe a considerable debt of gratitude to many people with regard to this book. First and foremost, I want to thank my mentors, Prof. Tianwen Zhang, Prof. Chenglin Liu, and Prof. Huizhi Lv. I could never have done this without their guidance and encouragement. I especially want to acknowledge the support of Prof. Fumitaka

Kimura who contributed the foreword for this book. Kimura's distinguished accomplishments, such as MQDF, have significantly shaped the field of handwritten Chinese character recognition. Special thanks go to Celine Chang and her assistant Jane Li at Springer Asia for their fantastic work in improving this book considerably. I would like to acknowledge my production editor Nithiya Sivaraman at Springer, and Mr. Srinivas and his typesetting team at Scientific Publishing Services, Chennai, India for the production of this book. I am sincerely grateful to Dr. Bin Lin and Dejun Guan for their constant encouragement, valuable feedback, and unconditional support throughout the long years of my academic career. I also want to thank the many colleagues and former students, who contributed their insights and invaluable comments. In pursuing my work, the following funds should be acknowledged: the National Natural Science Foundation of China (Grants no. 60475011, 60933010, 60975020 and 61203260), the Heilongjiang Natural Science Foundation of China (Grant no. F0322), and China Postdoctoral Science Foundation (Grant no. 20090450623). Finally, my special thanks go to my wife Sara Gong for her unfailing love, support, and patience.

31 October 2012 Tonghua Su

# Contents


**1 Introduction** . . . . . 1
1.1 Background . . . . . 1
1.2 Motivation . . . . . 2
1.3 Chronological Survey. . . . . 4
1.3.1 Infant Period (∼1970) . . . . . 4
1.3.2 Foundation Period (1970s) . . . . . 5
1.3.3 Exploring Period (1980s) . . . . . 6
1.3.4 Transcending Period (1990–2005). . . . . 7
1.3.5 Regenerative Period (Since 2006) . . . . . 9
1.4 Evaluation Metrics. . . . . 11
1.4.1 Character Segmentation Rates . . . . . 11
1.4.2 Recognition Rates . . . . . 12
1.5 Organization . . . . . 12
References . . . . . 14

**2 HIT-MW Database** . . . . . 23
2.1 Introduction . . . . . 23
2.2 Sampling Strategy . . . . . 26
2.2.1 Writer Sampling . . . . . 26
2.2.2 Text Sampling . . . . . 27
2.3 Database Acquisition . . . . . 27
2.3.1 Text Partition . . . . . 27
2.3.2 Layout Design . . . . . 28
2.3.3 Form Collection . . . . . 30
2.4 Database Processing. . . . . 30
2.4.1 Handwriting Digitalization . . . . . 30
2.4.2 Image Preprocessing . . . . . 30
2.4.3 Database Labeling. . . . . 31
2.4.4 Text Line Extraction . . . . . 32

2.5 Database Statistics . . . . . 37
2.5.1 Basic Statistics . . . . . 39
2.5.2 Erasing Statistics . . . . . 41
2.5.3 Miswriting Statistics . . . . . 43
2.6 Application of HIT-MW Database . . . . . 44
2.7 Discussions and Conclusions . . . . . 45
References . . . . . 46

**3 Integrated Segmentation-Recognition Strategy** . . . . . 49
3.1 Introduction . . . . . 49
3.2 Reference Systems . . . . . 50
3.2.1 Segmentation-Based Baseline Systems . . . . . 50
3.2.2 Integrated Segmentation-Recognition Baseline System . . . . . 53
3.3 Perceptron Learning of MQDF . . . . . 54
3.3.1 Perceptron Learning . . . . . 55
3.3.2 Dynamic Margin Regularization . . . . . 57
3.3.3 Interpretation . . . . . 58
3.4 Active Set Technique . . . . . 59
3.5 String Sample Synthesizing . . . . . 60
3.5.1 String Sample Distortion . . . . . 61
3.5.2 Verisimilitude Consideration . . . . . 63
3.6 String-Level Training . . . . . 63
3.6.1 MCE Criterion . . . . . 64
3.6.2 String-Level Training . . . . . 64
3.7 Linguistic Constraints . . . . . 65
3.8 Experiments . . . . . 66
3.8.1 Databases . . . . . 66
3.8.2 Evaluation on Character Databases . . . . . 67
3.8.3 Evaluation on HIT-MW Database . . . . . 73
3.9 Conclusions . . . . . 76
References . . . . . 77

**4 Segmentation-Free Strategy: Basic Algorithms** . . . . . 79
4.1 Introduction . . . . . 79
4.2 System Overview . . . . . 80
4.3 Sliding Windowing . . . . . 80
4.4 Feature Representation . . . . . 82
4.5 HMM Training and Decoding . . . . . 83
4.6 Experiments . . . . . 86
4.6.1 Evaluation on Segmentation-Free Systems . . . . . 87
4.6.2 Comparison Between Segmentation-Based and Segmentation-Free Systems . . . . . 93
4.6.3 Discussions . . . . . 95

4.7 Conclusions . . . . . 96
References . . . . . 97

**5 Segmentation-Free Strategy: Advanced Algorithms** . . . . . 99
5.1 Introduction . . . . . 99
5.2 Ensemble . . . . . 100
5.2.1 Curve of Character Matching Rate . . . . . 100
5.2.2 Serial Structure . . . . . 101
5.2.3 Parallel Structure . . . . . 102
5.3 Feature Enhancement . . . . . 104
5.3.1 Enhanced FPF (enFPF) . . . . . 104
5.3.2 Delta Features . . . . . 106
5.4 Text Line Normalization . . . . . 106
5.5 Synthesizing Samples . . . . . 106
5.6 Distributed N-gram Training . . . . . 109
5.7 MPE/MWE Training . . . . . 110
5.8 Experiments . . . . . 111
5.8.1 HIT-SW Database . . . . . 111
5.8.2 Evaluation of enFPF . . . . . 112
5.8.3 Evaluation of Ensemble Classifiers . . . . . 114
5.8.4 Evaluation of Text Line Normalization . . . . . 117
5.8.5 Evaluation of Delta Features . . . . . 118
5.8.6 Evaluation of Synthetic Samples . . . . . 118
5.8.7 Evaluation of MPE/MWE Training . . . . . 120
5.8.8 Evaluation of Distributed N-gram Training . . . . . 122
5.9 Conclusions . . . . . 122
References . . . . . 123