APR Advances in Pattern Recognition

Digital Document Processing

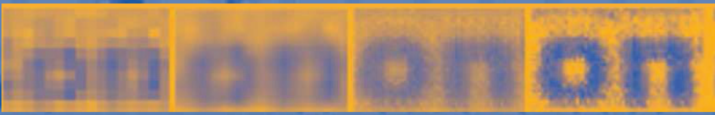
Bidyut B. Chaudhuri
(Ed.)

Major Directions and Recent Advances

Springer

APR Advances in Pattern Recognition

Digital Document Processing



Bidyut B. Chaudhuri
(Ed.)

Major Directions and Recent Advances

Springer

Advances in Pattern Recognition

Advance in Pattern Recognition is a series of books which brings together current developments in all areas of this multi-disciplinary topic. It covers both theoretical and applied aspects of pattern recognition, and provides texts for students and senior researchers.

Springer also publishes a related journal, **Pattern Analysis and Applications**. For more details see: http://link.springer.de

The book series and journal are both edited by Professor Sameer Singh of Exeter University, UK.

Also in this series:

Principles of Visual Information Retrieval
Michael S. Lew (Ed.)
1-85233-381-2

Statistical and Neural Classifiers: An Integrated Approach to Design
Šarūnas Raudys
1-85233-297-2

Advanced Algorithmic Approaches to Medical Image Segmentation
Jasjit Suri, Kamaledin Setarehdan and Sameer Singh (Eds)
1-85233-389-8

NETLAB: Algorithms for Pattern Recognition
Ian T. Nabney
1-85233-440-1

Object Recognition: Fundamentals and Case Studies
M. Bennamoun and G.J. Mamic
1-85233-398-7

Computer Vision Beyond the Visible Spectrum
Bir Bhanu and Ioannis Pavlidis (Eds)
1-85233-604-8

Hexagonal Image Processing: A Practical Approach
Lee Middleton and Jayanthi Sivaswamy
1-85233-914-4

Support Vector Machines for Pattern Classification
Shigeo Abe
1-85233-929-2

Bidyut B. Chaudhuri (Ed.)

Digital Document Processing

Major Directions and Recent Advances

Springer

Bidyut B. Chaudhuri, PhD
Indian Statistical Institute, Kolkata, India

Series editor
Professor Sameer Singh, PhD
Department of Computer Science, University of Exeter, Exeter, EX4 4PT, UK

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

Library of Congress Control Number: 2006927723

Advances in Pattern Recognition Series ISSN 1617-7916
ISBN-10: 1-84628-501-1 e-ISBN-10: 1-84628-501-1
ISBN-13: 978-1-84628-501-1 e-ISBN-13: 978-1-84628-501-1
Printed on acid-free paper.

© Springer-Verlag London Limited 2007

Apart from any fair dealing for the purposes of research or private study, or criticism or review, as permitted under the Copyright, Designs and Patents Act 1988, this publication may only be reproduced, stored or transmitted, in any form or by any means, with the prior permission in writing of the publishers, or in the case of reprographic reproduction in accordance with the terms of licences issued by the Copyright Licensing Agency. Enquiries concerning reproduction outside those terms should be sent to the publishers.

The use of registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant laws and regulation and therefore free for general use.

The publisher makes no representation, express or implied, with regard to the accuracy of the information contained in this book and cannot accept any legal responsibility or liability for any errors or omissions that may be made.

9 8 7 6 5 4 3 2 1

springer.com

Preface

The field of automatic document processing, more than a century old, is quite a mature one. It started with an attempt to automatically read printed text in the era well before the birth of digital computers and has been since continuing on various topics like document image enhancement, document structure and layout analysis, handwritten character recognition, document data compression, graphics recognition, document information retrieval and meta-data extraction. In addition to OCR, applications like tabular form and bank cheque processing or postal mail reading are of great interest to the software industry.

This edited book is a compilation of twenty chapters written by leading experts on the above and several other important topics. These chapters describe the state of the art in these areas, enumerate the research challenges and propose one or more possible solutions in a systematic way. Usually, edited books are compiled on some special area of a general discipline. But this one attempts to cover wider aspects of digital document processing and hence has the flavour of a handbook. Since there is no standard textbook with a wide coverage of the subject, this book will immensely help students taking undergraduate and graduate courses in digital document processing. Also, it is hoped that the researchers in the field will benefit from this book. About 9 years ago, Bunke and Wang edited a useful book of similar nature. However, the activities on document analysis have since advanced much, with newer techniques being invented by the community and younger disciplines like super-resolution text processing, handwriting individuality identification or web document mining gaining importance. Even on older topics like OCR, good work is in progress on challenging problems like the reading of Indian and Tibetan scripts. All these advancements contributed to the need for another edited book.

This book starts with an excellent introduction to the general discipline of document processing, followed by studies on document structure analysis. The next few chapters describe OCR systems for some difficult printed scripts that are followed by advances in on-line and off-line handwritten text recognition, with applications to postal automation and bank cheque processing. Then come the special topics of mathematical expression and graphics recognition, as well as super-resolution text analysis. Other problems like image degradation modelling, meta-data extraction, document information retrieval are addressed in the next few chapters. Some emphasis has been given on web document analysis and data mining problems. The last three chapters of this book are dedicated to these topics.

This book is the outcome of my interaction with the authors over a reasonably long period. In this endeavour, my sincerest thanks go to Professor Horst Bunke with whom I initially made plans to co-edit this book. At the preparatory stage, I received generous help from his vast experience on various issues like framing of book structure, the choice of chapter topics, the choice of potential authors as well as fixing other subtle points. However, because of heavy workload and subsequent illness, he had to discontinue his involvement in the editorial process. Nevertheless, I am grateful for his continuous advice, encouragement and best wishes for the completion of this project.

I express my sincere thanks to all authors for their hard work in preparing their manuscripts. Special mention should be made of Prof. L. Schomaker who not only wrote an introductory chapter but also patiently read all other chapters in order to write critical summaries of them. The readers will get a quick idea of all the topics by reading this chapter alone.

During my work on this book I enjoyed a sabbatical leave as well as a *Jawaharlal Nehru Fellowship*. The support of my institute and of the Jawaharlal Nehru Memorial Fund is gratefully acknowledged. I am thankful to my colleagues, Dr Utpal Garain, Mr Chittaranjan Das and Ms Shamita Ghosh, who helped me in various stages of the editorial work. The understanding and support of my family is also highly appreciated. Finally, I thankfully acknowledge the patience of Ms Catherine Brett of Springer Verlag for replying to my numerous e-mails and other support provided to me during the editing of this book.

Bidyut B. Chaudhuri
Kolkata, India

Contents

Preface v

Contributors xvii

1 Reading Systems: An Introduction to Digital Document Processing 1
Lambert Schomaker
1.1 Introduction 1
1.2 Text Sensing 3
1.3 Sensor Scope 3
1.4 Sensor Grid 6
1.5 Pre-processing 6
1.6 Invariance to Affine Transforms 7
1.7 Invariance to Ink-Trace Thickness 9
1.8 Shape Features 10
1.9 Processing Type 12
1.10 Computing Architecture 13
1.11 Computing Strategy 13
1.12 Knowledge Base 14
1.13 Cognitive Reliability 15
1.14 Response in Case of Difficult Input 15
1.15 Classification Accuracy 16
1.16 Energy and Mental Concentration 17
1.17 Processing Speed 17
1.18 Volume Processing 17
1.19 Summary of Human Versus Machine Reading 18
1.20 Conclusion 26
References 26

2 Document Structure and Layout Analysis 29
Anoop M. Namboodiri and Anil K. Jain
2.1 Introduction 29
2.1.1 Physical Layout and Logical Structure 30

2.2 Pre-processing 31
2.2.1 Noise Removal 31
2.2.2 Foreground Detection 32
2.2.3 Skew Correction 34
2.3 Representing Document Structure and Layout 34
2.4 Document Layout Analysis 36
2.4.1 Approaches to Layout Analysis 38
2.5 Understanding Document Structure 42
2.6 Performance Evaluation 43
2.7 Handwritten Document Analysis 45
2.7.1 On-Line Documents 45
2.8 Summary 46
References 47

3 OCR Technologies for Machine Printed and Hand Printed Japanese Text 49
Fumitaka Kimura
3.1 Introduction 49
3.2 Pre-Processing 49
3.2.1 Text Line Segmentation and Skew Correction 49
3.2.2 Character Segmentation of Japanese and English Mixed Text 51
3.2.3 Nonlinear Normalization of Character Size 53
3.2.4 Slant Correction 54
3.2.5 Segmentation–Recognition of Handwritten Japanese Document 56
3.3 Feature Extraction 58
3.3.1 Chain Code Histogram 59
3.3.2 Gradient Feature [16] 60
3.4 Classification 61
3.4.1 Linear Classifiers for High-Speed Pre-Classification 61
3.4.2 Quadratic Classifiers 62
3.5 Dimension Reduction 63
3.5.1 Discriminant Analysis 63
3.5.2 Principal Component Analysis 64
3.6 Performance Evaluation of OCR Technologies 64
3.6.1 Used Database 64
3.6.2 Dimension Reduction 64
3.6.3 Comparison on Pre-Classification 66
3.6.4 Comparison on Main Classification 67
3.7 Learning Algorithms 67
3.7.1 GLVQ 68
3.7.2 Performance Evaluation 69
3.8 Conclusion 69
References 70

4 Multi-Font Printed Tibetan OCR 73
Xiaoqing Ding and Hua Wang
4.1 Introduction 73
4.2 Properties of Tibetan Characters and Scripts 74
4.3 Isolated Tibetan Character Recognition 78
4.3.1 Character Normalization 79
4.3.2 Feature Formulation 81
4.3.3 Classifier Design 84
4.4 Tibetan Document Segmentation 88
4.4.1 Text Line Separation 88
4.4.2 Character Segmentation 91
4.5 Experiment Results 94
4.5.1 Performance of Isolated Character Recognition Algorithm 94
4.5.2 Performance of Segmentation Algorithm 95
4.6 Summary 96
References 96

5 On OCR of a Printed Indian Script 99
Bidyut B. Chaudhuri
5.1 Introduction 99
5.2 Origin and Properties of Indian Scripts 100
5.3 Document Pre-Processing 104
5.3.1 Binarization 104
5.3.2 Skew Detection/Correction 104
5.3.3 Text Region Detection 105
5.3.4 Text Line and Word Identification 105
5.3.5 Zone Separation and Character Segmentation 106
5.4 Character Recognition 107
5.4.1 Recognition of Upper Zone Shapes 107
5.4.2 Middle and Lower Zone Character Recognition 108
5.4.3 Punctuation Mark Recognition 111
5.4.4 Recognition of Touching Characters [9] 111
5.4.5 Combination of Results and Post-Processing 113
5.5 Performance Analysis 114
5.6 Conclusion 117
References 118

6 A Bayesian Network Approach for On-line Handwriting Recognition 121
Sung-Jung Cho and Jin Hyung Kim
6.1 Introduction 121
6.2 Modelling of Character Components and Their Relationships 124
6.2.1 Bayesian Network Modelling Framework 124
6.2.2 Point Model 125

6.2.3 Stroke Model 126
6.2.4 Character Model 128
6.3 Recognition and Training Algorithms 130
6.3.1 Recognition Algorithm 130
6.3.2 Training Algorithm 131
6.4 Experimental Results and Analysis 132
6.4.1 Data Set 133
6.4.2 Analysis of Within-Stroke Relationship Modelling . . 133
6.4.3 Analysis of Inter-Stroke Relationship Modelling 134
6.4.4 Comparison to Other Recognition Systems 136
6.4.5 Evaluation of Modelling Capacity by Generation of Characters 137
6.5 Conclusions 139
References 140

7 New Advances and New Challenges in On-Line Handwriting Recognition and Electronic Ink Management 143
Eric Anquetil and Guy Lorette
7.1 Introduction 143
7.2 On-Line Handwriting Recognition Systems 144
7.2.1 Main Features 144
7.2.2 General System Architecture 144
7.3 New Trends in On-Line Handwriting Recognition 144
7.3.1 Classical vs. New Acquisition Devices 144
7.3.2 Modelling 146
7.3.3 Classification 146
7.3.4 System Architecture 146
7.3.5 Linguistic Post-Processing 147
7.3.6 User Adaptation 147
7.3.7 From Off-Line to On-Line and from On-Line to Off-Line Handwriting Recognition 148
7.4 New Trends in Electronic Ink Management Systems 148
7.4.1 Pen-Based Applications 148
7.4.2 Digital Ink Annotations of Documents 149
7.4.3 Interpretation of On-Line Structured Documents . . . 150
7.4.4 Design of Handwriting Input for Small-Size Devices 154
7.5 Conclusion, Open Problems and New Challenges 156
References 157

8 Off-Line Roman Cursive Handwriting Recognition 165
Horst Bunke and Tamás Varga
8.1 Introduction 165
8.2 Methodology 166
8.2.1 Document Image Pre-processing 166

8.2.2 Recognition of Isolated Characters . . . 167
8.2.3 Cursive Word Recognition . . . 168
8.2.4 Cursive Word Sequence Recognition . . . 170
8.3 Emerging Topics . . . 171
8.3.1 Databases and Performance Evaluation . . . 171
8.3.2 Synthetic Training Data . . . 172
8.3.3 Multiple Classifier Systems . . . 174
8.4 Outlook and Conclusions . . . 175
References . . . 176

9 Robustness Design of Industrial Strength Recognition Systems . . . 185
Hiromichi Fujisawa
9.1 Characterization of Robustness . . . 185
9.2 Complex Recognition System: Postal Address Recognition . . . 187
9.3 Performance Influencing Factors . . . 189
9.4 Robustness Design Principles . . . 194
9.5 Robustness Strategy for Implementation . . . 203
9.6 Conclusions . . . 209
References . . . 210

10 Arabic Cheque Processing System: Issues and Future Trends . . . 213
M. Cheriet, Y. Al-Ohali, N.E. Ayat, and C.Y. Suen
10.1 Introduction . . . 213
10.2 Datasets . . . 214
10.3 Legal Amount Processing . . . 215
10.3.1 Pre-Processing . . . 215
10.3.2 Word Versus Sub-Word Processing . . . 216
10.3.3 Sub-Word Processing . . . 216
10.3.4 Legal Amount Interpretation . . . 221
10.4 Courtesy Amount Processing . . . 222
10.4.1 Pre-Processing . . . 223
10.4.2 Features Extraction and Representation . . . 224
10.4.3 Experiments . . . 226
10.4.4 Discussion . . . 229
10.5 Conclusion and Future Perspective . . . 230
References . . . 232

11 OCR of Printed Mathematical Expressions . . . 235
Utpal Garain and Bidyut B. Chaudhuri
11.1 Introduction . . . 235
11.2 Identification of Expressions in Document Images . . . 237
11.2.1 Techniques for Identification of Expression Zones . . . 237
11.2.2 Ways to Improve Identification Results . . . 239

11.3 Recognition of Expression Symbols 241
11.3.1 Existing Methods for Symbol Recognition 241
11.3.2 Ways to Improve Symbol Recognition Accuracy 242
11.4 Interpretation of Expression Structure 245
11.4.1 Previous Methods on Interpretation of Expression Symbols 246
11.4.2 Further Research on Interpretation of Expression Structure 247
11.5 Performance Evaluation 251
11.5.1 Basic Requirements for Performance Evaluation 252
11.5.2 Performance Evaluation Strategy 254
11.6 Conclusion and Future Research 255
References 256

12 The State of the Art of Document Image Degradation Modelling 261
Henry S. Baird
12.1 Introduction 261
12.2 Document Image Degradations 262
12.3 The Measurement of Image Quality 264
12.4 Document Image Degradation Models 266
12.4.1 Methodological Issues 266
12.4.2 A Physics-Based Model 267
12.4.3 A Statistics-Based Model 268
12.4.4 Estimation of Model Parameters 268
12.4.5 Estimation Using Images of Test Targets 268
12.4.6 Estimation Using Images of Text 269
12.4.7 Validation of Models 270
12.5 Applications of Models 270
12.5.1 Conditions for Connectivity and So Forth 271
12.5.2 Effects on Recognition Accuracy 271
12.5.3 Constructing Classifiers 271
12.5.4 Testing OCR Systems 272
12.6 Public-Domain Software and Image Databases 272
12.6.1 Simulation Studies 273
12.7 Open Problems 273
12.7.1 Uses of Synthetic Data 274
References 275

13 Advances in Graphics Recognition 281
Josep Lladós
13.1 Introduction 281
13.2 Application Scenarios 284
13.3 Early Processing 287
13.4 Symbol Recognition and Indexing 288
13.5 Architectures and Meta-data Modelling 289

13.6 On-Line Graphics Recognition and Sketching Interfaces 291
13.7 Performance Evaluation 293
13.8 An Application Scenario: Interpretation of Architectural Sketches 294
13.9 Conclusions: Sketching the Future 295
References 297

14 An Introduction to Super-Resolution Text 305
Céline Mancas-Thillou and Majid Mirmehdi
14.1 Introduction 305
14.2 Super-Resolution: An Analytical Model 307
14.3 MISO Super-Resolution: A Closer Look 308
14.3.1 Motion Estimation and Registration 308
14.3.2 Warping and Reconstruction 311
14.3.3 Regularization Techniques 315
14.3.4 Deblurring and Denoizing 317
14.3.5 Colour Super-Resolution Text 317
14.4 Case Study: SURETEXT – Camera-Based SR Text 318
14.4.1 Motion Estimation Using the Taylor Series 319
14.4.2 Unsharp Masking Using the Teager Filter 320
14.4.3 Outlier Frame Removal 321
14.4.4 Median Denoizing 322
14.4.5 Experiments and Results 323
14.5 Conclusions 325
References 325

15 Meta-Data Extraction from Bibliographic Documents for the Digital Library 329
A. Belaïd and D. Besagni
15.1 Introduction 329
15.2 The Users' Needs 330
15.3 Bibliographic Elements as Descriptive Meta-Data 331
15.4 Meta-Data Extraction in Bibliographic Documents 333
15.5 General Overview of the Work 333
15.6 Bibliographic Element Recognition for Library Management 335
15.6.1 Catalogue Retro-Conversion 335
15.7 Bibliographic Reference Structure in Technological Watch 341
15.7.1 Object Document 341
15.7.2 Target Document 343
15.7.3 Retrieval Engine 343
15.8 Citation Analysis in Research Piloting and Evaluation 344
15.8.1 Impact Factors 344
15.8.2 Methodology 345