Document Recognition V
Vol. 3305

PROCEEDINGS OF SPIE
SPIE—The International Society for Optical Engineering

Document Recognition V

Daniel P. Lopresti
Jiangying Zhou
Chairs/Editors

28–29 January 1998
San Jose, California

Sponsored by
SPIE—The International Society for Optical Engineering
IS&T—The Society for Imaging Science and Technology

Published by
SPIE—The International Society for Optical

SPIE
P
PROCEEDINGS SERIES
Volume 3305

Please use the following format to cite material from this book:
Author(s), "Title of paper," in *Document Recognition V*, Daniel P. Lopresti, Jiangying Zhou, Editors, Proceedings of SPIE Vol. 3305, page numbers (1998).

ISSN 0277-786X
ISBN 0-8194-2745-4

Published by
SPIE—The International Society for Optical Engineering
P.O. Box 10, Bellingham, Washington 98227-0010 USA
Telephone 360/676-3290 (Pacific Time) • Fax 360/647-1445

Printed in the United States of America.

Conference Committee

Conference Chairs

Daniel P. Lopresti, Lucent Technologies Bell Laboratories
Jiangying Zhou, Panasonic Technologies, Inc.

Session Chairs

1 Optical Character Recognition Techniques
Jonathan J. Hull, Ricoh California Research Center

2 Formulas, Maps, Tables, etc.
Kazem Taghva, University of Nevada/Las Vegas

3 Non-Roman Language OCR
Larry Spitz, Document Recognition Technologies, Inc.

4 Models and Evaluation
David S. Doermann, University of Maryland/College Park

5 Documents and the WWW
Luc M. Vincent, Xerox Corporation

6 Document Retrieval Systems and Issues
Michael D. Garris, National Institute of Standards and Technology

7 Document Image Processing
Ihsin T. Phillips, Seattle University

Introduction

The fifth annual Document Recognition conference consisted of 26 contributed papers and two invited talks, spanning nearly two full days of Electronic Imaging '98. As in previous years, the quality of the work presented at the conference was high. This is reflected in the papers included in this volume.

Participation in the conference illustrates the international nature of our field. In terms of papers, nine different countries are represented. The breakdown is:

11	from the United States
5	from Japan
2	each from China, France, and Germany
1	each from Australia, Finland, India, and the United Kingdom.

Our attendees came from industrial research labs, universities, and government agencies. All share a keen interest in document recognition and closely related problem areas.

We were fortunate to have two invited presentations to start off the conference. The first, by Larry Spitz of Document Recognition Technologies, was on applications of character shape coding. The second, by George Mills, concerned his work on architectures for segmenting hand-printed text for the Apple Newton.

The topics and structures of the Document Recognition conference will continue to evolve, with several important changes planned for 1999.

Daniel P. Lopresti
Jiangying Zhou

Contents

SESSION 1

Optical Character Recognition Techniques

Detection of Deleted Patterns in Handwritten Digits Using Topological and Geometrical Image Features

Misako Suwa, Satoshi Naoi, and Yoshinobu Hotta

Fujitsu Labs. LTD. , 4-1-1 Kamikodanaka, Nakahara-ku, Kawasaki-shi, Kanagawa-ken 211-88, Japan

ABSTRACT

One of the critical problems of an off-line handwritten character reader system is determining which patterns to read and which to ignore, as a form or a document contains not only characters but also spots and deletions. As long as they don't fit conditions for rejection, they cause recognition errors. Particularly, patterns of deleted single-character are difficult to be distinguished from a character, because their sizes are almost the same as that of a character and their shapes have variety.

In this article, we propose a method to detect such deletions in handwritten digits using topological and geometrical image-features suitable for detecting them; Eular number, pixel density, number of endpoint, maximum crossing counts and number of peaks of histogram. For precise detection, thresholds of the image features are adaptively selected according to their recognition results.

Keywords: handwritten character recognition, OCR, deletion, deletion detection

1. INTRODUCTION

Conventional OCR system cannot make a precise decision of patterns which must be read or must not be read and attempts to recognize all patterns in character fields as characters, as long as they don't fit conditions for rejection. Therefore, spots or deletions sometimes cause recognition errors. Especially, patterns of deleted single-character are much more difficult to be distinguished from a character than small dots or patterns of deleted multiple-characters. Because in the latter case, they can be removed or rejected judging from their sizes or sometimes peaks of their pixel-number histograms. But in the former case, the size of the pattern s are almost the same as that of an ordinary character and they have a variety of shapes. An example of an order form is shown in Fig.1. A symbol "?" indicates a rejected pattern. Some patterns of deleted single-character are misrecognized as digits and examples of patterns of deleted multiple-characters and deleted single-character are shown in Fig. 2 and Fig.3, respectively.

To decrease recognition errors, a function to distinguish a non-character from a character is needed in handwriting recognition. In this article, we propose a method to detect the patterns of deleted single-character in handwritten digits using topological and geometrical image-features.

Further author information -
M.S.(correspondence): Email: suwasan@flab.fujitsu.co.jp; Telephone: +81-44-754-2678; Fax: +81-44-754-2792
S.N.: Email: nao@flab.fujitsu.co.jp; Telephone: +81-44-754-2678; Fax: +81-44-754-2792
Y.H.: Email: hotty@flab.fujitsu.co.jp; Telephone: +81-44-754-2678; Fax: +81-44-754-2792

SPIE Vol. 3305 • 0277-786X/98/$10.00

In section 2, we will classify the patterns of deleted single-character according to their shapes. Image features suitable for quantifying the topological and geometrical characteristics of each type of deletions are suggested in section 3. Section 4 describes our deletion detection method, and the experimental results for evaluating the accuracy of the method is presented in section5. Section 6 is for conclusions and remarks.

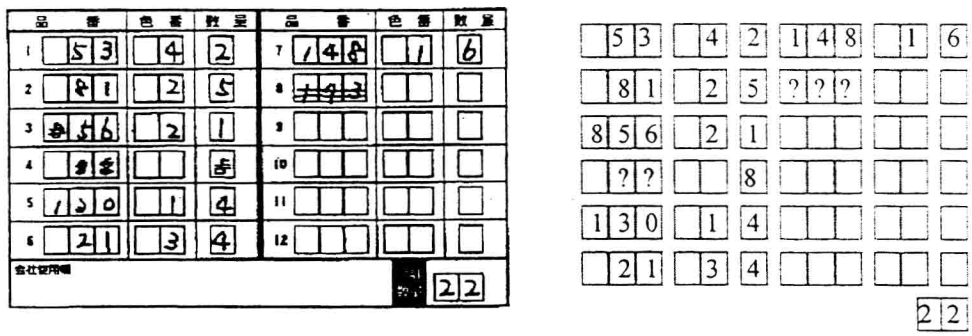


Fig. 1. A sample order form and its OCR recognition result

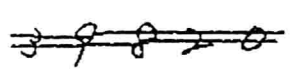


Fig. 2. Pattern of multiple-character deletion

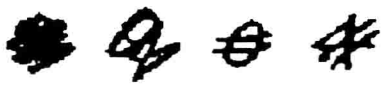
Fig. 3. Patterns of single-character deletion misrecognized as "8"

2. CLASSIFICATION OF SINGLE-DIGIT DELETIONS

First, we obtained 4000 samples of deleted patterns from 40 untrained writers and classified them roughly according to their shapes. This resulted in the four types shown in Fig. 4 and Table 1.

- Type a: A digit is painted out.
- Type b: Complex curves are drawn over a digit.
- Type c: Horizontal lines cross out a digit.
- Type d: Other types of lines - crosses, slashes, etc. - are drawn on a digit

In general, 'Type a' has a higher pixel density, 'Type b' has more loops, and 'Type c' and 'Type d' have more strokes, as compared with digits. 'Type c' also has peaks in its vertical-scanned histogram. Type c is the most common and it accounts for half of all deletions.



Type a Type b type c type d

Fig. 4. Four types of single-character deletion

Type	Type a	Type b	Type c	Type d
Ratio	3.0 %	14.9 %	47.3 %	34.8 %

Table 1. Ratio of each type of deleted single-character

3. SUITABLE IMGE FEATURES FOR THE DETECTION

To quantify the topological and geometrical characteristics of each type of deletions, the following image-features are thought to be suitable; pixel density, Eular number, number of endpoints, maximum number of crossing counts (stroke density), and number of peaks in a horizontally scanned pixel-number histogram..

(1) Pixel density

We define pixel density D as,

$$D = Sb / S,$$

where Sb is number of black pixels of the pattern and S is number of total pixels of its circumscribed rectangle. If a pattern is painted out, D becomes larger than that of an ordinary digit. To avoid mistaking a straight "1" for a deletion, we also use the ratio of width to height of the pattern,

$$R = \text{width} / \text{height}.$$

The R of "1" is expected to be larger than that of a deletion. So if D is larger than its threshold and R is less than its threshold, then we decide the pattern is a deletion.

(2) Eular number

Eular number E for thinned binary pattern[1] is known as follows,

$$E = C - L,$$

where C is number of connected components and L is number of loops of the pattern. Number of connected component of a handwritten digit is usually one or two, so if a pattern has more than two loops, E become minus. If E is less than its threshold, then we decide the pattern is a deletion.

(3) Number of endpoints

Number of end points P are also gained from a thinned pattern of the image[2], counting black pixels which have only one neighbor in their 8-neighborhood. If the number of line segments increase, P also increases. So if P is larger than its threshold, then we decide the pattern is a deletion.

(4) Maximum crossing counts

Horizontal and vertical crossing counts (stroke density) [3] are defined ad the number of strokes crossed by a horizontal and vertical scanned lines, respectively. We use maximum number of each crossing count; Ch and Cv. If the number of line segments increase, Ch and Cv also increase. So if Ch or Cv is larger than its threshold, then we decide the pattern is a deletion.

(5) Peaks in horizontal scanned pixel-number histogram

We also count the peaks in horizontal scanned pixel-number histogram H. If a pattern have horizontal line segments. the histogram has peaks. So if H is larger than its threshold, then we decide the pattern is a deletion.

4. DELETION DETECTION METHOD

We detect the deletions using difference of the features mentioned above between the deleted pattern and an ordinary digit. Of course, the ideal values of the features are also different among the 10 categories of digits. For example, handwritten "8" has at most two loops, but "6" has at most only one loop. So we set thresholds of the features for each category. We first pre-recognize an unknown pattern. According to the results of pre-recognition, a set of thresholds corresponding to the first candidate category is selected. To avoid being regarded a scraped pattern as a deletion with endpoints, we count the number of connected components of the pattern. If its connected components are larger than , we regard it as a scraped pattern and
remove it from the detection process.

The deletion detection algorithm runs as follows;

(1) An unknown binary pattern is pre-recognized.

(2) The number of connected components of the pattern is calculated. If it exceeds the threshold, we regard the pattern as a scraped pattern and stop processing.

(3) Each image-feature − pixel density, Eular number, number of endpoints, maximum crossing counts and number of peaks in the pixel number histogram − is extracted from the pattern.

(4) The thresholds of the features corresponding to the category of recognition result are set adaptively.

(5) Using the thresholds, we decide whether the pattern is a deletion or an ordinary digit.

The simple flow of the process is illustrated in Fig.5.

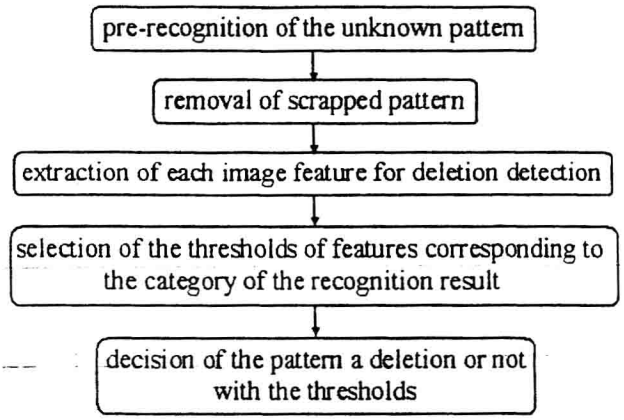


Fig. 5. Algorithm flow

5. EXPERIMENTAL RESULTS

To establish thresholds, we collected 4000 samples of deletions and 4000 samples of ordinary digits from 40 untrained writers. We also collected 2000 samples of deletions and 2000 samples of ordinary digits for testing. The directional code feature with Quasi-Mahalanobis distance[4] was used for character recognition. We decided the thresholds for detection error rate for ordinary digits to be less than 0.1%. The results are shown in Table 2.

On the training set, the algorithm could detect about 76% of deleted patterns and the detection error for ordinary digits was 0.08%. On the test set, a detection rate of 75% of deleted patterns was achieved and the detection error for ordinary digits was 0.05%. In Table 3, we give a breakdown of detection with each feature. Some deletions were multiply detected. Examples of detected deletions for each feature are shown in Fig.6. The detection errors for ordinary digits were caused mainly by extreme deformation of a digit or spots on a digit as you can see from Fig.7. Examples of deleted patterns which cannot be detected by this method is also shown in Fig.8. Some deletions have quite resembled ordinary digits in their shapes. And also deletions whose segments overlap with strokes were hard to detected.

An Example of the recognition result by the conventional OCR without the deletion-detection and with the deletion-detection are shown in Fig.9. As you can see from Fig. 9 (b), we could reject the deleted patterns and reduce the recognition errors in Fig. 9 (a).

Data set	Detection rate for deletions	Error rate for ordinary digits
Training	76.5%	0.08%
Test	75.1%	0.05%

Table 2. Detection rate for deleted patterns and error rate for ordinary digits

Pixel density	Eular number	Endpoint	Crossing counts	Histogram
171	1518	1247	848	1083

Table 3. Breakdown of detection

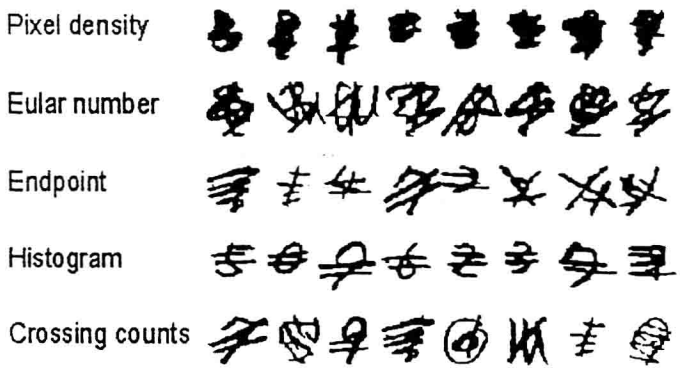


Fig. 6. Examples of detected patterns for each feature

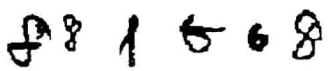


Fig. 7. Examples of detection error for digits

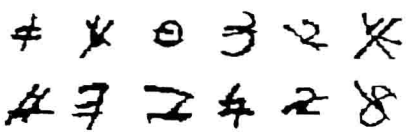


Fig. 8. Examples of detection omission for deleted patterns

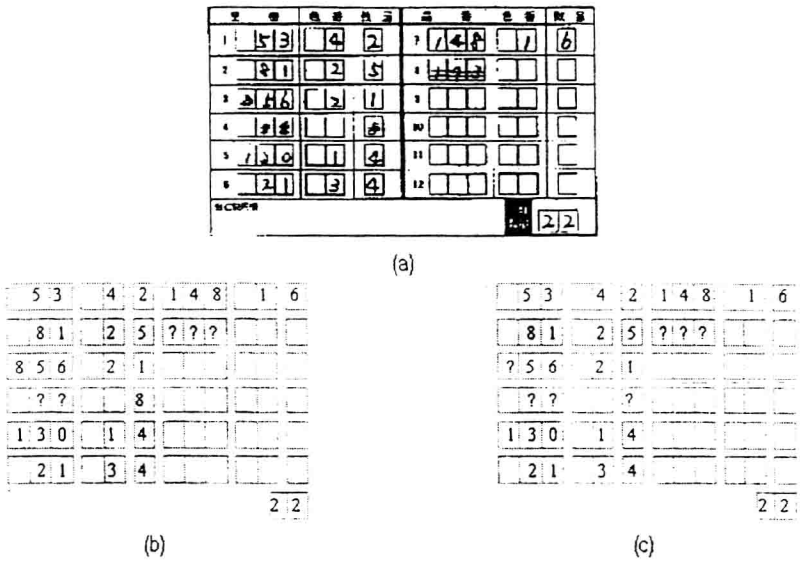


Fig. 9. An Example of recognition result with deletion detection and without deletion detection (a) Original image (b) Recognition by conventional OCR (c) Recognition by conventional OCR + deletion detection

6. CONCLUSIONS AND REMARKS

Using topological and geometrical image-features we have proposed an algorithm to detect patterns of deleted single-digit in handwritten digits. Pre-recognition results - the first candidate category - were used to select thresholds of the features. With this algorithm, a detection rate of 75% was achieved and the detection error rate for ordinary digits was 0.05%. Combing this algorithm with a conventional character recognition method, we could reject recognition errors of the deletions.

In this time we treat only handwritten digits, but our final goal is to realize a function to distinguish all non-characters from all characters in an OCR system.

7. ACKNOWLEDGEMENTS

We wish to thank Jun'ichi Hirai who has given us assistance in programming this algorithm.

8. REFFERENCES

1. S. Yokoi et al, "Topological Properties in Digiti*CE Japan*, **Vol.56-D**, No.11, pp662-669, 1973

2. H. Bunke and P. S. P. Wang, "Handbook of Character Recognition and Document Image Analysis", *World Scientific*, 1997

3. S. Naito et al, "Stroke Density Feature for Handprinted Chinese Characters Recognition", *IEICE Japan*, **Vol.J64-D**, No.8, pp757-764, 1981

4. F. Kimura et al., "Modified Quadratic Discriminant Functions and the Application to Chinese Character Recognition", *IEEE Trans. Pattern Anal. Machine Intell.*, **Vol.PAMI-9**, pp149-153, 1987.