Line Removal and Restoration of Handwritten Characters on the Form Documents

Jin-Yong Yoo, Min-Ki Kim, Sang Yong Han and Young-Bin Kwon
Computer Vision Lab., Dept. of Computer Science and Engineering
Chung-Ang Univ., 221 Heuksuk-Dong, Dongjak-Ku,
Seoul 156-756, Korea
e-mail:{yong, mkkim, ybkwon}@ripe.chungang.ac.kr, han@peace.cse.cau.ac.kr

Abstract

The off-line handwritten characters recorded on prescribed form documents may be overwritten by the lines of form documents. Overwitten characters should be isolated in order to be more effectively recognized. However, removal of the line causes breaks in overwritten characters. Consequently, a character restoration process is necessary. In this paper, the shape types of overwritten characters are analyzed and a method of restoring characters broken by line removal is proposed. A 97% correct restoration ratio was obtained through this method.

Keyword : form document, line removal, character restoration, handwritten character.

1 Introduction

Form documents of prescibed shapes are daily used at public institutes, government and public offices. One phenomena which occurs when writting on documents is characters overwritten by the lines of form documents. To be precise, a character recognizer extracts segmented characters by removing the form lines of overwitten characters. In this process, a procedure to join or restore broken characters due to removed lines is required[1].

Recognition of handwritten characters is considered very difficult work. If a recognition program is processed with overwritten characters, a high recognition rate cannot be expected, no matter how excellent the recognizer used. The reason for the poor recognition rate is geometrical series increasement of the subject of the recognition owing to excessive transformation[2]. Therefore, line removal and character restoration is required for recognition of a bill, a receipt slip or a paying-out slip with the prescribed form. One possible way of line removal and character recognition is a method of mathematical morphology[3]. The line of a check with the overwritten English characters is removed by using opening and closing operation. This method is easy to employ. On the other hand, it has a weak point in that the broken characters are not correctly restored[4]. Another method is proposed by Srihari[5]. In some algorithms, the broken characters are restored. If result that character recognizer is performed with the restored characters is wrong, restored characters are sent back to the restoration algorithm stage. In this method, the processing time is increased because it has feedback paths. In addition, characters are sometimes recognized incorrectly such as 'h' and 'b'. Thus, we analyze the shape of overwritten Hangul(Korean characters) from the prescribed forms. Then, we classify the type of shapes touching lines.

In this paper, we propose an algorithm which takes the line removal and the restoration of the overwritten characters on the line. This paper consists of 4 parts : analysis, implementation, experimentation and conclusion. In the analysis stage, results of overwritten characters obtained from the 616 different writers are described. Based on the type analysis, a restoration procedure which includes a line removal algorithm is implemented. Following the experimental results, conclusions are presented.

2 Feature Analysis of Overwritten Characters

2.1 Methods of Analysis

A precise analysis is a prerequisite to correct restoration of characters. Namely, it analyzes plenty of collected data and classifies general cases. We made a prescibed form and distributed it to the students of Chungang university and their family members. We attempt to analysze the overwritten patterns of 616 sheets in two steps. At first, the overwritten part of collected data is divided into two parts whether overwritten line is horizontal or vertical. Each part is then divided into a straight line and a curved line of the overwritten characters. It is sometimes ambiguous to distinguish the difference between a straight line and a curved line. Moreover, subdivision of each element is also difficult. Secondly, we classify the collected data into a number of junction point. The arrows of Fig 2.1 indicate the examples of junction point. The junction points are rearranged by the structural shape of contact part afterwards. A junction point is defined as a contact point of characters with line. In other words, after data are categorized as a number of junction point, same junction point numbers are classified

0-8186-7898-4/97 $10.00

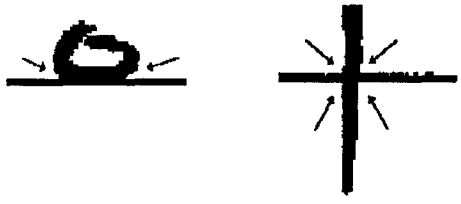

Fig 2.1 Line and junction points

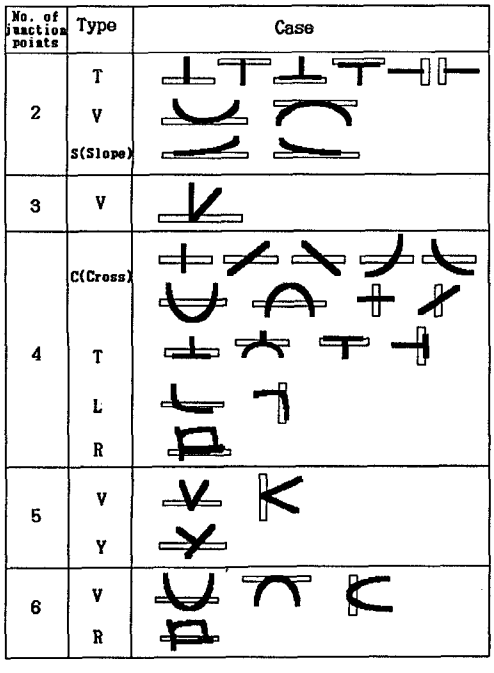

Fig 2.2 Classification of types based on the junction points and contacting shapes

by shape. The name of subdivided type is called by a similar capital English letter as shown in Fig 2.2. In the figure, black line stands for the handwritten character and white line stands for the prescribed line.

2.2 Contacting Shape Analysis

From the analysis on the collected data, an overlapping between line and characters is classified with 5 different classes based on the number of junction points. Then, crossing shapes can be categorized into 7 different types (that is, T, V, S, L, Y, R, and C type) from their contacting forms. Table 1 shows total 13 different classes based on the junction points and contacting shapes. From now, we define a type name using a combination of junction points and contacting shapes such as 2/T.

Table 1: Modified classes

Results of analysis		Modified type
No. of junction points	Type	
2	T V S	simple line removal
3	V	4/V
4	C T L S R	4/C 4/T 4/S 4/S 4/T
5	V Y	6/V 6/V
6	V R	6/V 6/R

3 Implementation of a Line Removal and Restoration Algorithm

3.1 Overview of a proposed algorithm

An algorithm which consists of a line extraction, a line removal, a storage of junction points, a decision of type and a restoration of broken characters is proposed. In the implementation of a proposed algorithm, a line of a form documents is not only a correct straight line but also a irregular line with various thickness. Structural shapes of the classified type is also partially duplicated. Therefore, a reduction of type is possible in order to implement a proposed algorithm. For example, in Fig 2.2 simple line removal is sufficient if the junction point is two. After a line removal, slight deformation of the character is made. But the character itself keeps its original shape. The other junction point numbers keep the six defined types. Thus, an implementation of the line removal and restoration is considered only when the junction points are greater than two. After considering the similar shape analysis, we can reduce the shapes to the similar one with different junction points. This modification is summarized in Table 1. Using this modified classes, we can develop a proposed line removal and restoration algorithm.

3.2 Explanation of a proposed algorithm

3.2.1 Line extraction procedure

Input data characters may simultaneously overlap both the top line and the bottom line. We assume that there are two form lines. Because major objective of the research is line removal and restoration of broken characters, portions of characters descending from the

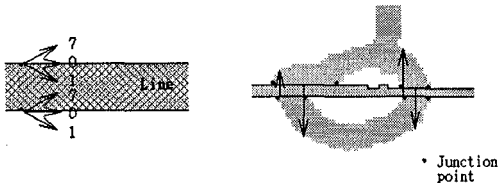

Fig 3.2 Line tracing and junction points decision

above line need not be removed. It is assumed that input data is correctly scanned, that is, skew angle is zero. If a horizontal black run is longer than a certain length threshold, we assume it is a line.

3.2.2 Line removal and junction detection

We trace the top and bottom of the black run. Fig 3.2 shows that three-direction(1,0,7) are checked and traced[6]. First, the top and bottom of the black run are traced only in the 0 direction. If a black pixel is in the 1 or 7 directions, it is checked whether another black pixel exists above or below that point. If such a point exists, the algorithm removes the line in the 0 direction. The algorithm also searches in the 1 or 7 direction for a junction point until it finds a black pixel. If the distance between the top and the bottom of the black run exceeds the threshold, the point is stored as a contacting point.

3.2.3 Restoration of broken characters

In the case of detected junction points are two, a line is merely removed as described in Table 1. The rest of types are 6/V, 6/R, 4/C, 4/T, 4/S, and 4/V type as a modified classes of Table 1. The 6/V type is V type that the number of junction point is 6. It is an inverse trapezoidal shape that junction points of the top of a line's black run is four and junction points of the bottom of a line's black run is two. In this case, first and fourth junction points of upper 4-junction point are connected with the two bottom's junction points. The R type is also a trapezoidal shape that the number of junction point is six. Thus, restoration is done as same as the 6/V type. The 4/C type is a cross shape that the number of junction points is four. In this case, corresponding junction points of the top and bottom black runs are connected with a line along the horizontal or vertical crossing. 4/T type is a T shape or an inverse of the T shape with the number of junction points is four. Corresponding junction points of the top and bottom black runs are connected with a line. The error of this connection is relatively tiny on the narrow width of line. A 4/S type is a slightly tilted shape that the number of junction point is four. Corresponding junction point is connected along the line. The 4/V type has four junction points on the top black run only. Restoration of this type is accomplished by connecting the first and the fourth junction points on the top of black run and virtually two junction points of a bottom line.

Table 2: Statistics of experimental data

Type	The number of type	%
Cross type	63	43.16
2 junction points	40	27.40
Slope type	18	12.33
4/T type	12	8.22
6/V type	7	4.79
4/V type	5	3.42
R type	1	0.68

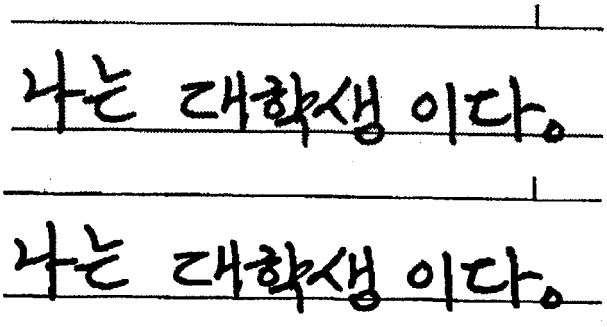

Fig 4.1 Gray scale and binary image

4 Experimental Results

The line removal and restoration of characters were done on a Axil 311 workstation using the C language. The sampled data was scanned as 400-dpi gray scale images by a AGFA ARCUS scanner. There were no limitations placed upon the size of the characters. Thus, overall or part of item which was overwritten on the form lines was scanned. For example, overall address was scanned or a part of address was scanned in which the handwritten data is overlapped with prescribed line. One sampled data is shown in Fig 4.1.

Scanned data also includes two form lines. The data is converted into a binary image. The number of occurances determined by the program of each of the various types of junctions between characters and form lines is giving in Table 2.

A total of 111 characters and 23 numerals were collected. Among them, 83 characters and 5 numerals overlapped a form line. Out of 146 junction points, 141 were correctly restored (96.58%) and five were incorrectly restored.

Fig 4.1 shows a grayscale and binary image. Fig 4.2 shows the image after line removal and the resulting restoration is depicted in Fig 4.3. The white gaps created during line removal and clearly restored.

In some cases, noise or cursive variation of characters shape resulted in incorrect restoration. As Fig 4.4 demonstrates, some minor errors may occur. The character as a whole, however, retains the characteristics Hangul; it looks like a noisy character. With

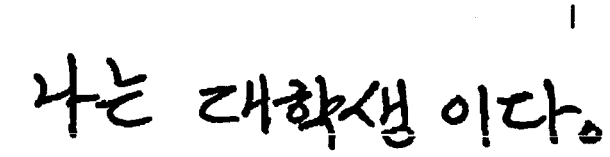

Fig 4.2 Line removal

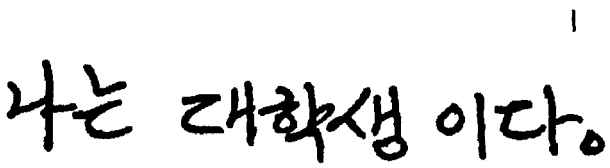

Fig 4.3 Restored image

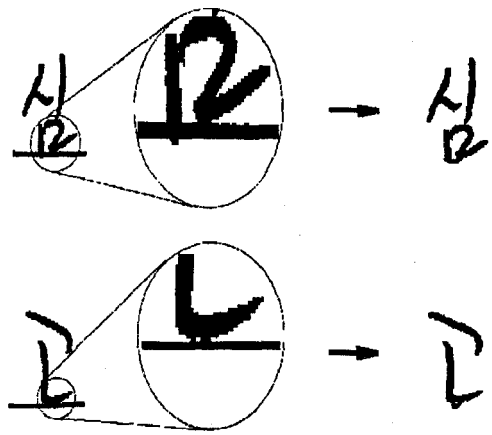

Fig 4.4 Examples of incorrect restoration

that in mind, we believe that even incorrect restoration works well enough to enable the recognition of handwritten characters using a simple noise filtering.

5 Conclusions

We proposed a line removal and a character restoration method which are based on the morphological analysis of the contacting shape and the number of junction points. After analyzing collected data of handwritten Korean characters, we classified the data into six types. We received a 97% correct restoration ratio from our implemented algorithm. In the future, it will be necessary to restore on vetical line as well as horizontal lines on credit card slips. From the error analysis, we determined that an effective method of line extraction and an alaysis method to elaborate the restored data are needed. The extension into the English alphabet is also considered.

Acknowledgement

This work was supported by KOSEF(93-0100-02-01-3) and Chungang University.

References

[1] Ying Liu, Richard French, Sargur N. Srihari, "An Object Attribute Thresholding Algorithm for Document Image Binarization", *International Conference on Document Analysus and Recognition,* pp. 278-281, 1993.

[2] Shunji Mori, Ching Y. Suen, and Kazuhiko Yamamoto, "Historical Review of OCR Research and Development", *Proc. of the IEEE,* Vol. 80, No. 7, pp. 1029-1058, July 1992.

[3] Charles R. Giardina, Edward R. Dougherty, *Morphological Methods in Image and Signal Processing,* Prentice Hall, Inc., 1988.

[4] Didier Guillevic, Ching Y. Suen, "Cursive Script Recognition: A fasr reader scheme", *International Conference on Document Analysis and Recognition,* pp 311-314, 1993.

[5] D. Wang, S. N. Srihari, "Analysis of Form Images", *International Conference on Document Analysis and Recognition,* pp. 181-186, 1991.

[6] H. Freeman, J. M. Glass, "Computer processing of line drawing images", *ACM Computing Surveys,* Vol. 6, No. 1, pp. 57-97, 1974.