A Segmentation Method for Touching Handwritten Japanese Characters

Hiromitsu Nishimura[1], Hisashi Ikeda[2], and Yasuaki Nakano[1]

[1] Dept. of Information Engineering, Shinshu University, Nagano, Japan
[2] Central Research Laboratory, Hitachi Ltd., Kokubunji, Japan

Abstract. The purpose of this paper is the segmentation of touching handwritten Japanese characters to enable each segmented character to be recognized. Though segmentation methods cooperating with recognition for the simple characters are known, they are not applicable to complicated characters such as kanji. To estimate the ability of segmentation itself, our method does not use any results of character recognition but segments touching characters using some specific features of character patterns. Linear components in patterns are extracted as features and a segmentation method based on them is proposed. A database, which contains character pattern and ground truth is constructed. In the database the ground truth data describe the correct areas of touching points judged by plural human subjects. As a result of the segmentation experiments, about 64% of touching patterns can be segmented at appropriate points. This correct segmentation rate is automatically calculated consulting the ground truths.

1 Introduction

The demands for the recognition of freely handwritten Japanese characters have become stronger, so as to recognize addresses of letters, manuscripts, memos and so on. In these applications, touching handwritten characters become big problems.

In European languages, words are commonly written in cursive style, so there are character recognition methods (called holistic approach) which identify a pattern as a word without segmenting touching characters.

On the other hand, in Japanese, there is no space between words and the number of combinations of neighboring characters is too big, so that the recognition of touching handwritten characters may be almost impossible if they are not segmented.

For handwritten alphanumeric characters, methods using over-segmentation and multiple hypotheses cooperating with recognition have been reported [1]. For cursive handwritten English word recognition, a method that uses recognition results and lexicon matching has also been reported [3]. For printed word recognition, a segmentation method that uses character sizes has been reported.

S.-W. Lee and Y. Nakano (Eds.): DAS'98, LNCS 1655, pp. 130–139, 1999.

In the recognition of handwritten Japanese address characters a method cooperating segmentation, recognition and lexical matching was reported [5]. But the methods coping with touching characters have not been studied deeply.

In this paper, the authors propose a segmentation method that does not use any recognition results, but only uses specific features of patterns. The reason is in that the recognition performance of handwritten kanji characters is not thought high enough to verify the multiple hypotheses.

Using a newly built document database with ground truth data of touching areas, a series of the segmentation experiments was done. The correct segmentation rate was automatically assessed.

1.1 Specific Features of Handwritten Japanese Character Scripts

Handwritten Japanese character scripts have the specific features as shown in Figure 1.

a: Both vertical (top to bottom) and horizontal (left to right) writing directions exist.
b: Curved pattern characters (mainly hiragana and Arabic numerals) and linear pattern characters (kanji and katakana) are used.
c: Character sizes are not uniform.
  c1: Character sizes of numerals are much smaller than kanji
  c2: Kanji characters are usually bigger than hiragana characters
d: Kanji characters are complicated and constituted of many strokes.

In this paper, we deal with not general handwritten Japanese characters but only handwritten Japanese address characters.

The segmentation methods [2] [3] [4] using “Contour following analysis”, “Character size”, “Peripheral distribution”, which are proposed for the recognition of English handwritten address characters, are not appropriate, because of the specific features of Japanese characters stated above.

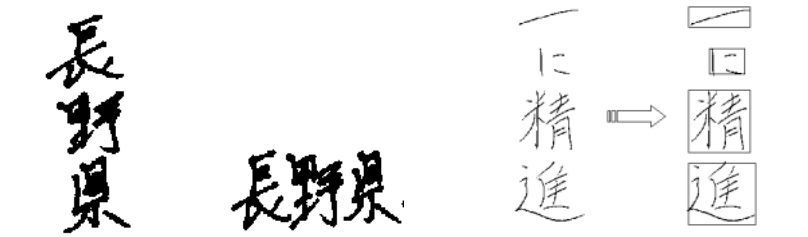

Fig. 1. Examples of handwritten Japanese character script

1.2 Proposed Method

As stated in the condition b in the previous section, there are kanji, hiragana, katakana and alphanumerical characters in handwritten Japanese scripts. If we restrict ourselves on the addresses on the mails, however, most of the characters are kanji. Kanji are mainly constructed with linear components. Thus most of touching characters are found between kanji. In the many touching cases, the last stroke of the preceding character extends too long and touches to the one of strokes of the next character.

In this paper, we assume that a kanji touches with another kanji. Under the assumption, a method to estimate linear components and segment touching patterns is proposed.

2 Database

A Japanese address character database is supplied by Postal Administrator Laboratory for the research encouragement. Since it is not clear how many touching patterns are included in the database, another database fitted to the purpose of the research is needed.

By the reason, the database stated in the section 2.1 is prepared and used in this research.

2.1 The Database for Segmentation Experiments

Data collected for segmentation experiments consist of 8,000 images written by 10 subjects with ball-point pens and fiber-tipped pens in both vertical and horizontal styles. At the collection, the subjects were not instructed to write touching characters intentionally. By forcing the subjects write patterns into rather smaller regions, however, the touching characters were generated in natural manners.

After digitizing the 8,000 images with the resolution of 200 dpi, each of them was examined by human subjects if it includes touching characters. Thus a collection consisting of 2,487 touching patterns selected from the 8,000 images was constructed. (All collected images can be recognized and segmented by human subject.)

2.2 Ground Truth Database

In the researches reported so far, some promising ideas have been proposed, but the experimental verifications have been shown in a few examples. In this sense, the assessments to guarantee the results seem to have been lacking.

In this paper, results are not tested in any character recognition system. So the progresses in character recognition rate can not evaluate the propriety of segmentations.

In this paper, ground truth data judged by human subjects have been made in order to assess the results of the proposed method automatically.

To make the ground truth database, each original scanned pattern was displayed on the computer screens and four subjects were forced to pick the touching areas using the mouses. The point clicked as the touching one was converted to a small square region including the point. The results given by four subjects were merged and an area corresponding to the touching point was generated by a majority logic.

2.3 Possibilities of Recognition by Segmentation

A touching pattern can not be recognized in a usual Japanese character recognition system, because such a usual system is designed only to recognize each segmented character. If a touching pattern is segmented correctly, each segmented pattern may be considered as recognizable.

But not all the segmented patterns can be recognized even if the correct segmentation is done. The reason may be in that touching characters might be deformed heavily so as not to be identified after the segmentations. Besides, the recognition method may not be complete.

As a result, the correct segmentation rate will not completely agree with the recognition rate improvement. It is expected, however, that many correctly segmented patterns have high possibility to be recognized correctly, thus a higher recognition rate may be induced.

3 Segmentation Using Straight Line Components

In the steps explained below, the segmentation lines in the patterns are estimated.

As the precondition, the patterns to be segmented should be extracted from preprocessing stage. By extracting conected regions from an address image, some of them are judged to have too large widths or heights. They can be the touching regions of the two components belonging to the different characters. Each of such conected regions is supplied to the segmentation unit.

3.1 Estimation of the Linewidths and Pattern Normalization

Addresses in mails are written with various character sizes and with various kinds of pens. To make the proposed method robust, it is necessary to normalize the linewidths of patterns.

In the Figure 2, to estimate the linewidth of a pattern, the length in the direction perpendicular to every contour point independently in every small block is estimated. Then the frequency distribution along the stroke is calculated. From the maximum point of the frequency distribution the linewidth of every small block is estimated.

For patterns having too large linewidth, the proposed method may search many surplus linear components from the patterns. To avoid surplus searches, patterns having a large linewidth are preprocessed by thinning.

The thinning method is not the usual "thinning" which converts patterns to those having one pixel linewidth. Our method normalizes the patterns having too large linewidths to those having the standard width. In the following algorithm, two formulae are examined independently in vertical and horizontal directions. In the algorithm

m, n: the estimated and standard linewidths of a character,
X: abscissa, Y: ordinate, O: input image, N: normalized image,
O (x, y), N (x, y): current coordinates in images,
Value of the pixel: 1 for black pixel, 0 for white pixel.

Algorithm

$$Vertical\ thinning \begin{cases} if\ \mathrm{O(x,y+u)=1}\ for\ \mathrm{u=\{-(m\text{-}n),\ldots,(m\text{-}n)\}} \\ \qquad\qquad\qquad \mathrm{N\ (x,\ y)=1}, \\ else \qquad\qquad\quad \mathrm{N\ (x,\ y)=0}, \end{cases}$$

$$Horizontal\ thinning \begin{cases} if\ \mathrm{O(x+u,y)=1}\ for\ \mathrm{u=\{-(m\text{-}n),\ldots,(m\text{-}n)\}} \\ \qquad\qquad\qquad \mathrm{N\ (x,\ y)=1}, \\ else \qquad\qquad\quad \mathrm{N\ (x,\ y)=0}. \end{cases}$$

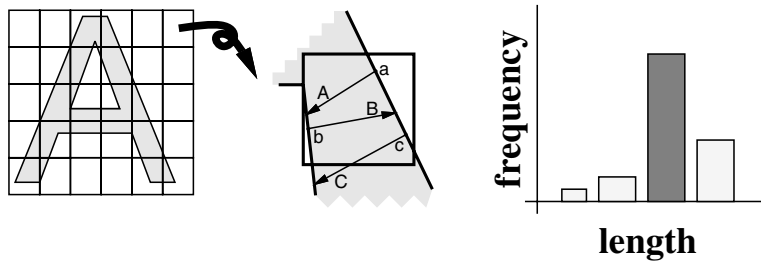

Fig. 2. Estimation of the linewidth of a written character

3.2 Extraction of Linear Components

Since the touching patterns to be segmented are mainly constructed of linear components, extraction of linear components is necessary. In this paper, we estimate linear components from Freeman-code strings extracted from contours of patterns.

As is well known, Freeman-code has only eight values of resolution of the directions that are rather poor. So, we adopt the following condition C to extract linear components in more subtle directions. After extractions of linear components, the extracted linear components are quantized into the eight directions.

If the following five conditions are satisfied for a part of the contour, it is extracted as a candidate of a linear component.
Algorithm

A: If same Freeman-code continues for more than a threshold, the continuation is estimated as a linear component.
B: If some other codes contained in the sequence of the same code and their number is less than a threshold, the continuation including the other codes is estimated as a linear component.
C: When A and B do not produce a string of a linear segment, averaging operation is tested. If the averaged value of some continuous codes is close to the value of the mode of the codes, the continuation is estimated as a linear component.
D: If the length of a linear components estimated by A, B and C is smaller than a threshold, the result is rejected.
E: Linear components are extended along the estimated direction. If the pixels on the extension line have the value "1", they are added to the linear components.

3.3 Generation of Simplified Patterns and Estimation of Potential Segmentation Areas

From a hypothesis that the potential segmentation areas in complicated patterns are to be estimated by a macroscopic viewpoint, simplified patterns are generated.

The simplified pattern is generated by overlapping the extracted linear components, thickening the overlapped area and dilating it. By using simplified patterns, many surplus segmentation lines are suppressed.

The segmentation areas are estimated by extracting crossing points of linear component from the simplified patterns.

If the area around the crossing point has more pixels in the overlapped area than a threshold, the square region around the crossing point with a prescribed edge is estimated as a potential segmentation area.

An example is shown in Figure 3, where the leftmost shows the original image, the center the simplified pattern and the rightmost the potential segmentation areas.

3.4 Estimation of Segmentation Lines

In the step explained below, segmentation lines in potential segmentation area estimated in section 3.3 is extracted.

A potential segmentation area is searched by the scanning line to the direction of character writing. The uppermost (leftmost) point and lowermost (rightmost) point is determined along the every scanning line. Tentative segmentation

Fig. 3. Estimation of potential segmentation areas

lines are estimated by linking the uppermost (leftmost) and lowermost (rightmost) points. So the number of the tentative segmentation is the number of the combination of those two points.

If a tentative segmentation line estimated above does not satisfy the following conditions, it is rejected.

1. The tentative segmentation line segments a conected region.
2. The numbers in pixels of the segmented areas exceed a threshold.

The shortest tentative segmentation line along the all estimated lines is decided as the segmentation line.

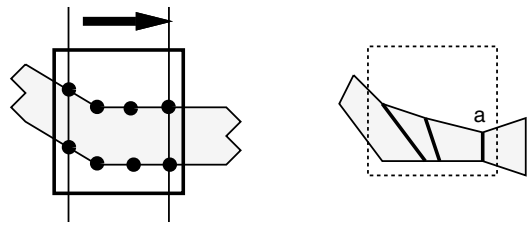

Fig. 4. Example of estimation of segmentation lines

3.5 Reduction of the Surplus Segmentation Lines and Segmentation

By the algorithm explained so far, potential lines for the segmentation are extracted. But many surplus segmentation lines are included in them. Since the combination of the image fragments after the segmentation is fed to the character recognition unit, the number of the combination should be minimum from the viewpoint of the recognition time. Therefore, surplus segmentation lines should be reduced.

The surplus segmentation lines are reduced by scoring. Each estimated segmentation line is scored by the following formula.

$(\Sigma(Pixels\ in\ segmented\ area)) \times (Pixels\ in\ original\ area)$

There are two reasons for adopting the scoring formula.

1. The larger continuous area is, the more probable it is to be segmented.
2. The closer a segmentation line to the center of a continuous area is, the more feasible it is.

Each segmentation result is assessed by the comparison with the ground truth data.

Table 1. The distribution of all estimated correct segmentation lines on score rank

$P = \frac{number\ of\ reduced\ segmentation\ lines}{number\ of\ all\ segmentation\ lines}$		
score rank	P	correct segmentation rate
3	16.3%	25.1%
5	25.8%	40.9%
7	33.1%	48.0%
9	38.2%	51.9%
∞	100%	64.3%

4 Experimental Results

4.1 Automatic Assessment

Automatic assessing system is constructed using the ground truth database stated in section 2.2. In the system, if a part of estimated segmentation lines is included in a ground truth data, the segmentation line is judged as the appropriate segmentation line. If any part of the line is not included in the ground truth area, the segmentation line is estimated as inappropriate.

4.2 Experiment and Assessment

Table 1 shows the comparison of the elimination rate of candidates and the correct segmentation rate as the function of candidate number. Between the result marked as "∞" and that marked as "9", it can be seen that by using the scoring the number of segmentation lines is reduced by 61.8% at the cost of 12.4% decrease of the detection rate. From the observation, the reduction method is considered to be effective.

It may be doubtful, however, that the automatic assessing system works exactly or not. To evaluate the correctness of the results, all segmentation results

are also judged by human subjects. The results are shown in Table 2. In Table 2, each result is judged by the following rules. If it seems correct, label it "correct". If it seems incorrect, but the segmented patterns seem to be recognizable, label it "recognizable". If no segmentation lines are detected, but the original image seems recognizable, label it "failed but recognizable". If no segmentation lines are detected or only surplus segmentation lines are detected (false alarm), label it "failure".

Namely, the former three cases are labeled as the segmentations are proper and the last case is labeled as the segmentations are false. Though there is a little difference in four subject answers, "correct segmentation" is about 60% which is very close to the automatic assessed result.

Table 2. Human assessment for segmentation results

	subject1	subject2	subject3	subject4
correct	66.4%	63.0%	60.3%	53.0%
recognizable	13.3(79.7)%	13.2(76.2)%	19.8(80.1)%	20.6(73.6)%
failed but recognizable	4.1(83.8)%	7.5(83.6)%	4.0(84.1)%	5.1(78.7)%
failure	16.2%	16.4%	15.9%	21.3%

(The percentage in the parentheses denotes the sum.)

5 Concluding Remarks

About a half of the tested data of the database was segmented correctly by the proposed system. The automatic assessing system judges correct answers strictly. Since the assessing is too strict, some segmentation lines close to the ground truth areas are classified as failures even if they may be able to segment the touching patterns so as to enable each segmented character to be recognized. To evaluate the possibility, we made four subjects assess the suitability of the results. From the results about 80% of the touching patterns are segmented properly.

Though the correct segmentation rate does not completely agree with recognition rate improvement as stated in section 2.3, it is highly possible that the recognition system rate will be improved.

61.8% of the surplus estimated segmentation lines are reduced by the proposed scoring method at the cost of 12.4% correct segmentation rate reduction as shown in Table 1.

Using the proposed method, many touching handwritten Japanese characters are segmented, without so many surplus segmentations, so as to enable each segmented character be recognized.

Acknowledgments

The authors appreciate many beneficial advices by Dr. Hiromichi Fujisawa and Dr. Hiroshi Sakou both at Central Research Laboratory, Hitachi Ltd. The authors appreciate the cooperation given by the members of Nakano-Maruyama Laboratory of Shinshu University.

References

1. H.Fujisawa, Y.Nakano and K.Kurino: Segmentation Methods for Character Recognition: From Segmentation to Document Structure Analysis. Proc. of IEEE. **80–7** (1992) 1079–1092 130
2. Hidefumi Ino, Kazuki Saruta, Nei Kato and Yoshiaki Nemoto: Handwritten Address Segmentation Algorithm Based on Stroke Information. ISSN. **38–2** 1997 131
3. Hirobumi Yamada and Yasuaki Nakano: Cursive Handwritten Word Recognition Using Multiple Segmentation Determined by Contour Analysis. IEICE. Trans. INF&SYSTE. **E79–D–5** (1996) 130, 131
4. Masaomi Nakajima and Yuji Yonekura: Handwritten Character Segmentation Using Smoothing Histgram and Discriminant Analysys. IEICE. Trans. INF&SYSTE. **J78–D–2–7** (1995) 131
5. Yayoi Kobayashi and Jun Tsukumo: Handwritten Characters Recognition Using Contexts. IEICE. Trans. **PRU91–67** (1991) 39–46 (In Japanese) 131