MVA '98 IAPR Workshop on Machine Vision Applications, Nov. 17-19, 1998, Makuhari, Chiba, Japan

13—3 Segmentation of Handwritten Kanji Numerals Integrating Peripheral Information by Bayesian Rule

Tatsuhiko Kagehiro†, Masashi Koga†, Hiroshi Sako†, Hiromichi Fujisawa†

Central Research Laboratory, Hitachi, Ltd.,

Abstract

We have developed a new segmentation method for Kanji numerals. Segmentation of Kanji numerals written vertically is difficult because part of one Kanji numeral pattern can be read as another Kanji numeral character. In our method, Kanji numerals are segmented correctly by using peripheral information and the similarity given by the character classifier. However, the usefulness of this information and the similarity differs depending on the character categories. Thus we represented the degree of usefulness by a likelihood ratio that is selected according to the character categories. The likelihood ratios are integrated ideally by using the Bayesian rule. When implementing this method, we achieved character segmentation accuracy of 77%, and address recognition accuracy of 63%. In both cases our obtained accuracy was much higher than with a conventional segmentation method.

1. Introduction

We have developed a new method of segmenting handwritten Kanji numerals written vertically. The segmentation of Kanji numerals written vertically is difficult, because part of one Kanji numeral pattern can be read as another Kanji numeral character. Figure: 1(a)-(d) show Kanji numerals and hyphens written vertically. In Figure: 1(a), there are three Kanji numerals and two hyphens, showing the Kanji numeral "three", hyphen, "two", hyphen, and "one" from top to bottom. The Kanji numeral "one" is expressed by one horizontal stroke, "two" is expressed by two horizontal strokes; and "three" is expressed by three horizontal strokes. Thus, if two horizontal strokes are written in a vertical direction, the pattern is ambiguous, and can be read as a pair of one's or a single two. This ambiguity is what makes Kanji numerals written in a vertical sequence difficult to segment.

There are three kinds of information that can generally be used to segment characters:[1]

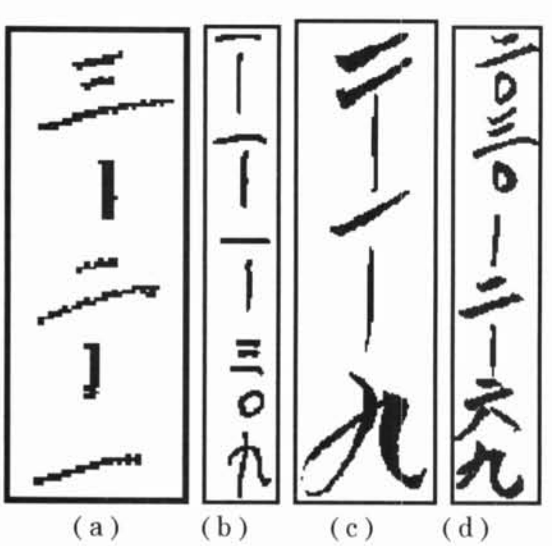
(1) peripheral information like size, shape, and relative positions of patterns[2],

(2) similarity of a candidate character pattern(s) to a valid character[3], and

(3) word matching score[4].

In the case of Kanji numeral segmentation, type (1) and (2) information can be used effectively, but type (3) information can not be used because unlike a string characters the numeral sequence has no contextual information. Type (1) information is effective for character segmentation but is not sufficient by itself. Type (2) information is particularly useful for valid segmentation, but it is risky to use it alone because segmentation in some cases is highly ambiguous as described above.

Our method uses both type (1) and (2) information to resolve the ambiguities. The first problem we faced was the fact that each of the peripheral features included in the type (1) information, for example width, height, aspect ratio, and gap, don't have same degree of usefulness for ambiguity resolution. And more, the degree of usefulness depends on the character



(a) (b) (c) (d)

Figure: 1 Kanji numerals written vertically
(a) 3-2-1, (b) 1-1-1-309, (c) 2-1-9, (d) 2030-2-69

† Address : 1-280 Higashi-koigakubo, Kokubunji-shi, Tokyo 185-0014, Japan
E-mail : kagehiro@crl.hitachi.co.jp

category. To overcome these problems, we applied the Bayesian rule to obtain confidence values for each potential reading. These confidence values enable us to increase the segmentation accuracy.

2. Conventional character segmentation method

Segmentation is done as follows[5]. An input image is analyzed by a pre-segmentation module that makes multiple candidates of segmented character patterns. In Figure: 2, there are pre-segmented patterns of the Kanji numeral "three" (三). The hypotheses of the pre-segmentation are represented in terms of a graph (or a network), and one of the paths from the initial node to the terminal node is the result of segmentation. To select the optimum path among all possible paths, the links between nodes which represent pre-segmentation candidates, are evaluated. The similarity of each pre-segmented pattern is used in the conventional segmentation method. The valid patterns (links) have a high similarity value, and the invalid patterns (links) have a low value. In Figure: 2,one of the pre-segmented patterns is inputted into the character classification module, then the character category "三" and similarity value of 0.9 are obtained. In this case, the similarity is high because the inputted pattern has a valid shape. The evaluation of the links between nodes eventually identifies the optimum path. This search for the optimum path can be done through dynamic programming.

However, many pre-segmented patterns of Kanji numerals could be a part of several Kanji numeral patterns. As shown in Figure: 2, the pre-segmented patterns for "三" consist of one horizontal line, two horizontal lines, and three horizontal lines. All patterns can be classified as the Kanji numerals "one", "two", or "three" with a high degree of similarity. Therefore, all

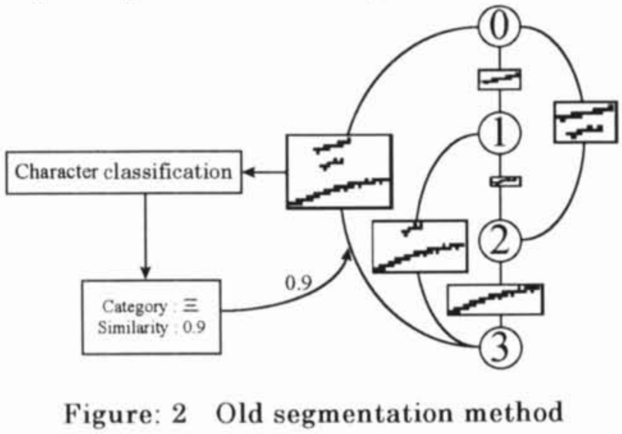


Figure: 2 Old segmentation method

evaluation values are very large, so it is difficult to determine the correct path from these hypotheses of the pre-segmentation in the case of Kanji numerals.

3. Information concerning character category and peripheral features

3.1. Character category information

When the peripheral features of Kanji numerals "three", hyphen, "two", hyphen, and "one" (Figure: 1(a)) are determined, the aspect ratio of the Kanji numeral "three" is about 1.0. In contrast the width of Kanji numeral "one" is much greater than the height. Also, the Kanji numeral "three" consists of three connected components, while the Kanji numeral "one" consists of one. In this way, the valid values of the peripheral features differ depending on the character categories.

How useful a feature is also differs depending on the character categories. For example, the usefulness of the spaces in front of and behind the characters is different from Kanji numerals "three" and "one" because the Kanji numeral "one" has the same shape as one part of the Kanji numeral "three", but a single horizontal line can be distinguished as either "one" or one part of Kanji numeral "three" by using information concerning the spaces. Therefore, for the Kanji numeral "one", the information concerning the spaces in front of and behind characters is very important, and the usefulness of this feature is high.

On the other hand, the similarity given by the character classifier is not important and is not useful for Kanji numeral "one" because the similarity values for "one" and the one part of "three" is large in both cases. However, this information is very useful for other character categories.

Thus, we have to determine the usefulness of specific features for each character category, and the degrees of usefulness can then be used to calculate the link evaluation values.

3.2. Usefulness of features

The degree of feature usefulness can be presented as a likelihood ratio that is given by,

$$L(e_k|Hc) = \frac{P(e_k|Hc)}{P(e_k|\overline{Hc})}$$

where

c : character category assumed for a candidate,
Hc : the positive hypothesis of the link being a pattern of character category c assumed for a candidate,
e_k : one of the measured features, or the evidence of the hypothesis.
$P(e_k / Hc)$: the probability of e_k supposing event Hc..

The likelihood ratio is then evaluated for the assumed category. We computed the conditional probability $P(e_k / Hc)$ in terms of the histogram of e_k when event Hc occurred. To make this histogram, features e_k are extracted from all pre-segmented patterns, then classified into the corresponding character categories, and further divided into two classes, namely the correct and incorrect segmentation classes. So $P(e_k / Hc)$ can then be obtained from the histogram. By using the above formula, likelihood ratio $L(e_k / Hc)$ can be obtained from this $P(e_k / Hc)$.

Figure: 3 shows the likelihood ratio graph of a feature extracted from the pre-segmented character patterns. This graph shows the likelihood distribution in the case of the character category of Kanji numeral "one", the horizontal axis is the feature value of space in front of and behind the character, and the vertical axis is the likelihood ratio. Increasing the space value increases the likelihood ratio, demonstrating that the feature of the space in front of and behind characters is useful.

Thus, after extracting many features from pre-segmented patterns, we selected the most useful features for character segmentation. These features are shown below.

(A) pattern height (normalized by line width)
(B) pattern width (normalized by line width)
(C) pattern aspect ratio
(D) spaces in front of and behind characters
(E) number of connected components in pattern
(F) similarity of character classifier

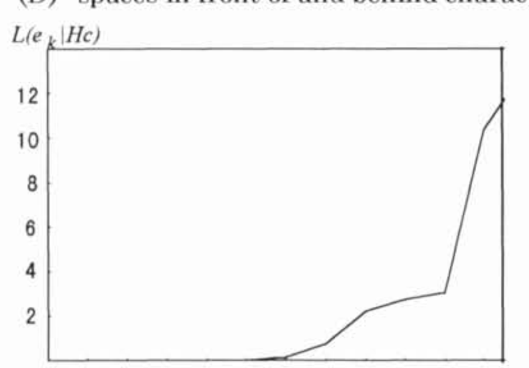


Figure: 3 Likelihood ratio graph

3.3. Integrate likelihood ratios

The method we used to calculate the confidence value of a link by using the likelihood ratios of features is described below. We define this confidence value as the *a posterior* probability $P(Hc / e_1, e_2, e_3, \ldots, e_n)$, where $\{e_1, e_2, e_3, \ldots, e_n\}$ is a list of measured features, or the evidence of the hypothesis. In this paper, $\{e_1, e_2, e_3, \ldots, e_n\}$ are the (A)-(F) features listed above. The confidence value can be transformed into a computable formula by applying the Bayesian rule (Bayes theorem), as shown below.

$$P(Hc|e_1, e_2, e_3, \ldots e_n) = \frac{\frac{P(Hc)}{P(\overline{Hc})} \prod_{k=1}^{n} L(e_k|Hc)}{1 + \frac{P(Hc)}{P(\overline{Hc})} \prod_{k=1}^{n} L(e_k|Hc)}$$

where
c : character category assumed for a candidate,
H_c : positive hypothesis of category c given the candidate,
$e_1, e_2, e_3, \ldots e_n$: measured features of the candidate, and
$L(e_k | H_c)$: likelihood ratio for e_k for correct category c.

The usefulness of each feature is represented by the likelihood ratio. Thus, the calculated confidence value can be used for the link evaluation.

4. Implementation

4.1. Learning stage

In the learning stage, many sample images are pre-segmented where candidates of the character patterns including over-segmented partial patterns are extracted. Peripheral features and similarity values are extracted from all of the pre-segmented patterns, and the histogram of these features is made as described in section 3.2. The sampled versions of the conditional density functions are then calculated, and the likelihood ratios $L(e_k / Hc)$ are pre-computed in a non-parametric way. The calculated $L(e_k / Hc)$ of each feature is stored and used at the recognition stage.

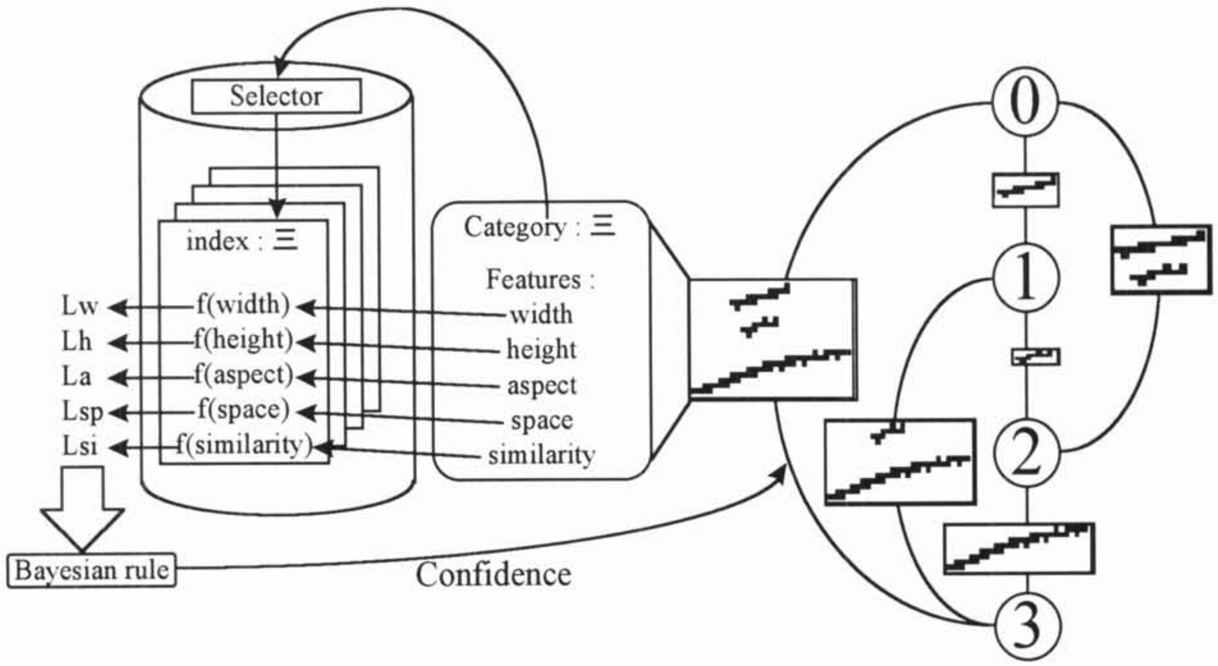


Figure: 4 New segmentation method

4.2. Recognition stage

When an image is to be recognized, the pre-segmentation module performs segmentation as described above, and the features for each candidate are gathered. Among the features, there is a category index c given by a character classifier, and the confidence is computed for the assumed category. The computation process is depicted in Figure: 4. Selection of the best path is then made using these confidence values.

5. Experimental results

We evaluated the proposed method by applying it to destination address recognition of handwritten Kanji letters. An experiment was done for each of two versions of the address recognition system, one using our proposed method and the other using conventional method. We used 253 sample images of handwritten Kanji numerals. In these samples, characters were written vertically with a ball-point pen, or a writing brush, or a fiber-tipped pen, and so on.

As shown as Table:1, the character segmentation accuracy increased by almost 50point, and the address recognition accuracy increased by about 35point, demonstrating the effectiveness of our method for handwritten Kanji numerals. However, this method is much less effective in a situation where some characters are connected, so another method that creates pre-segmented candidates from the connected characters is also needed.

Table:1 Experimental results

	Conventional	New method
Segmentation accuracy	29%	77%
Recognition accuracy	26%	63%

References

[1] R.G.Casey et al., "A Survey of Methods and Strategies in Character Segmentation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 18, No. 7, pp.690-706, July 1996

[2] N. Babaguchi, et al., "A Fundamental Study on Character Segmentation from Handwritten Japanese Character Strings," *IEICE (in Japanese)* Vol. J68-D, No. 12, pp.2123-2131, Dec. 1985

[3] H. Murase, et al., "Writer Adaptive Character Segmentation Method for Free Format Handwriting, " *IEICE (in Japanese)* Vol. J72-D-II No. 1, pp.132-139, Dec. 1985

[4] A. Suzuki, et al., "Recognition Method Handling Character Position Gaps in Handwritten Japanese Address," *IEICE (in Japanese)* Vol. J77-D-II No. 1, pp.20-28, Jan. 1994

[5] H. Fujisawa, et al., "Segmentation Method for Character Recognition," *in Proc. IEEE*, Vol. 80 No. 7, pp1079-1092. July 1992