Word Segmentation in Handwritten Korean Text Lines
Based on Gap Clustering Techniques

Soo H. Kim*, S. Jeong **, Guee-Sang Lee***, Ching Y. Suen*

*Center for Pattern Recognition and Machine Intelligence
Concordia University, Suite GM606, 1455 de Maisonneuve Blvd. West
Montreal, Quebec H3G 1M8, Canada

**Postal Technology Development Department, ETRI
161 Kajong-dong, Yusong-gu, Taejon 305-600, Korea

***Department of Computer Science, Chonnam National University
300 Yongbong-dong, Buk-gu, Kwangju 500-757, Korea

Abstract

We propose a word segmentation method for handwritten Korean text lines. It uses gap information to separate a text line into word units, where the gap is defined as a white-run obtained after a vertical projection of the line image. Each gap is classified into a between-word gap or a within-word gap using a clustering technique. We take up three gap metrics – BB, RLE and CH which are known to have superior performance in Roman-style word segmentation, and three clustering techniques – average linkage method, modified MAX method and sequential clustering. An experiment with 498 text line images extracted from live mail pieces has shown that the best performance is obtained by the sequential clustering technique using all three gap metrics.

I. Introduction

Many studies on recognizing off-line handwritten words or characters have been published. But most of them assume their input as isolated units. It is important to be able to recognize sentences or phrases in numerous applications: in the interpretation of handwritten addresses [1], tax forms [2], FAX messages [3], and so on. A typical approach to recognize a text line involves the division of a line into words and recognizing each of them. In this paper, we focus on the separation of a handwritten Korean text line into words by determining the location of gaps between words.

Although there is no published paper related to the segmentation of handwritten Korean text lines, several methods for Roman-style word segmentation have been proposed [4-6, 12]. These methods can be divided into two kinds according to the feature set used: one using the gap between connected components [5, 7-10], and the other adopting additional features which a human uses when he/she separates words in a line image [4, 12]. These features, for example, are presence of punctuation marks, presence of upper case characters followed by a string of lower case ones, transition between numerals and alphabetic letters, and so on. But it is nearly impossible to obtain this information without a recognition stage.

In this paper we use the gap information for word segmentation, utilizing the fact that the size of gaps between words is larger than that of gaps within a word. In this regard, word segmentation is composed of two stages. First, the gaps are identified from the text line and their magnitudes are estimated, and then the gaps are classified into between-word gap or within-word gap using a clustering technique.

A number of gap metrics have been proposed for the segmentation of Roman-style words [4, 5, 12], and we adapt the best three of them to see whether they are also effective for the segmentation of handwritten Korean words. In [4], Seni and Cohen explored eight algorithms that estimate the gap and evaluated their performance. In [5, 12], Mahadevan et al compared the performance of their gap metric with those in [4].

In the gap clustering stage, we consider three clustering techniques – average linkage method, modified MAX method and sequential clustering. The average linkage method is one of hierarchical agglomerative approaches and was first introduced for solving the problem of word segmentation in [12]. The MAX method in [12] is a kind
of threshold-based clustering technique, and we have modified this method to make it more effective for word segmentation. Finally, a new sequential clustering method is proposed in this paper.

Our goal is to find the best combination of gap metric and clustering technique for handwritten Korean word segmentation. We have observed through an experiment that the best performance of word segmentation is obtained when the sequential clustering technique using all the three gap metrics is applied.

The remainder of this paper is divided into four sections. Section 2 describes the three gap metrics in detail and Section 3 discusses how to classify the gaps. The experimental results and their discussion are given in Section 4 and finally a concluding remark is given in Section 5.

II. Measurement of gap distance

Gap metrics for Roman-style text lines regard the gap as the space between two connected components (CC’s), but we define the gap as the space between two overlapped components (OC’s), where an OC is defined as a set of CC’s whose projection profiles in the vertical direction are connected. This modification is based on the observations: (1) there are many CC’s in a Korean character and the space between the CC’s within a character is not useful for word segmentation; (2) the characters in Korean words are written with some space.

We detect a set of gaps by a vertical projection of the input line image, where the gap is a white-run in the projection profile. We assume that the skew of the input line image has been corrected, and there is no overlap between two words (assumption supported by the test data). Figure 1 shows two between-word gaps (g2, g3) and five within-word gaps (g1, g5, g4, g6, g7).

Figure 1. Gaps from a vertical projection of line image

2.1 Bounding box distance: BB

The BB measure has been proposed by Seni and Cohen [4] for the segmentation of an English text line into words. We adapt this measure by computing the distance between adjacent bounding boxes of OC’s. This measure is known to have a drawback in the segmentation of Roman-style words - the rectangular nature of the bounding box suppresses component shape, and hence this metric tends to underestimate the size of gaps [4, 5]. However, we choose it as an effective gap metric for Korean texts because each character is written in a rectangular shape. Figure 2 shows some gap distances computed using the BB measure.

![Figure 2. Gap distances by the BB measure (unit: pixel)](image)

2.2 Run-length/Euclidean distance: RLE

The RLE measure is also proposed by Seni and Cohen [4]. We measure the RLE distance by utilizing one of the minimum run-length or the minimum Euclidean distances between adjacent OC’s. The choice of one of the two is made by considering a heuristic rule with the degree of overlap.

The degree of overlap is calculated using the equation, O/(H1 + H2), where H1 and H2 indicate the heights of the left and right OC’s, and O indicates the magnitude of vertical overlap between the two OC’s. The minimum run-length is used if the degree of overlap is larger than a threshold; otherwise, the minimum Euclidean distance is chosen. We set 0.25 as the threshold. The two types of RLE distance are given in Figure 3.

![Figure 3. Minimum run-length/Euclidean distances](image)

Figure 4 presents some gap distances computed by the RLE measure. All distances are the minimum run-length because all the adjacent OC’s overlap more than the threshold. Note that the RLE measure always produces larger values than the BB distances.

![Figure 4. Gap distances by the RLE measure (unit: pixel)](image)

2.3 Convex hull distance: CH

The CH measure, proposed by Mahadevan et al. [5], can also be applied to measure the size of gaps between OC’s. First we compute the convex hull of each OC, which is the smallest convex polygon enclosing the component, using Graham’s algorithm [13]. Given a pair of adjacent components C1 and C2, let H1 and H2 be their convex hulls. Let L be the line joining the centers of gravity of H1 and
III. Gap clustering

The problem of word segmentation is to define a function $f$ that maps a gap in a line image to a between-word gap (BWG) or a within-word gap (WWG). If the set of gaps is \{g_1, g_2, ..., g_n\}, the function $f$ takes one of three forms: (1) it maps every $g_i$ to the WWG if the image contains only one word, (2) it maps every $g_i$ to the BWG if the image contains $n+1$ words, and (3) it maps some $g_i$'s to the WWG and the others to the BWG.

We use a heuristic rule to distinguish the first two types from the last. First, the gap sizes are normalized using their mean and standard deviation. Let $R$ be a difference between the maximum and minimum sizes among the normalized gaps. The heuristic rule detects the first two types of $f$ if $R$ is less than a threshold. Next, to distinguish the first from the second, the normalized mean is used. All the gaps are classified into the BWG if the mean is less than a threshold; otherwise all of them are classified into the WWG. The thresholds are set through an experiment. For the third case, we consider the following clustering techniques to determine the membership of each gap.

3.1 Average linkage method

The average linkage method is one of hierarchical agglomerative clustering techniques, and was first introduced to solve the problem of word segmentation in [12]. Initially, each gap is treated as a singleton cluster. In successive steps, the two closest clusters are merged into a larger cluster until all the gaps have been grouped into one cluster. We use the following equation to measure the distance between two clusters,

$$D(C_i, C_j) = \frac{1}{n_i n_j} \sum_{a \in C_i, b \in C_j} D_{ab}(a, b),$$

where $n_i$ (or $n_j$) is the number of gaps in a cluster $C_i$ (or $C_j$) and $D_{ab}$ indicates the function of Manhattan distance.

The result is a dendrogram like the one shown in Figure 6 which illustrates the process of merging seven clusters into one using the BB distances in Figure 2. It is organized in such a way that the right cluster of any node has a larger mean than its left cluster. Each hypothesis which is a possible set of BWG's is produced by choosing a node from the dendrogram and selecting all its in-order successors.

![Figure 6. Dendrogram of gaps and a list of hypotheses](image)

3.2 Modified MAX method

The MAX method stated in [12] is based on the following idea. In a sorted order of gaps, it is very likely that the maximum difference between consecutive gaps occurs between the lowest word gap and the highest non-word gap. With this observation, the MAX method finds this boundary and classifies all the gaps above the boundary as word gaps.

We modify the MAX method by considering also the ratio of magnitudes between consecutive gaps. Given a sorted list of gaps, $g(1), g(2), ..., g(n)$, we compute the difference $d_i$ between consecutive gaps and the ratio $r_i$ of their magnitudes as follows:

$$d_i = D_{i+1}(g(i), g(i+1))$$

$$r_i = \left| \frac{g(i)}{g(i+1)} \right|$$

where, $D_{i+1}$ is a function of Euclidean distance and $| \cdot |$ is calculated by the norm whereas $g(i)$ is a multi-dimensional vector. The boundary is determined at the point where the product of $d_i$ and $r_i$ is the maximum. The i-th ranked hypothesis is generated on the basis of the i-th maximum of the products. Figure 7 shows the process in which this method is applied to the data in Figure 2 and the resulting hypotheses.

![Figure 7. A ranked list of hypotheses produced by the modified MAX method](image)
3.3 Sequential clustering

In this paper, we propose a new clustering technique referred to sequential clustering. Given a sorted list of gaps, \(<g(1), g(2), ..., g(n)>\), this method starts with two clusters initialized as follows:

\[
C_{\text{WWG}} = \{g(0)\}, \quad C_{\text{BWG}} = \{g(n)\}
\]

where, \(g(0)\) is a null gap and indicates the smallest within-word gap. We set the value of \(g(0)\) to ‘0’. For each of the remaining \(n-1\) gaps, its membership is determined to belong to one of two clusters according to the distance between the gap and the centroids of current \(C_{\text{WWG}}\) and \(C_{\text{BWG}}\). The order of membership classification is as follows: \(g(1), g(n-1), g(2), \) and so on. The reason that we set the order like this is to reduce the bias of information that happens when the data continue to be put into only one cluster. The stop condition occurs when the smallest remaining gap is classified into \(C_{\text{BWG}}\) or the biggest remaining one is classified into \(C_{\text{WWG}}\).

Let the smallest gap in the final \(C_{\text{BWG}}\) be \(g(j)\). The i-th ranked hypothesis is generated using the following equation:

\[
C_{\text{BWG}} = \{g(j-K), ..., g(n)\}, \quad \text{where } K = j+i-1-n
\]

Figure 8 illustrates the process of generating the two clusters and shows the first three hypotheses. The 1st hypothesis is \(\{g(6), g(7)\}\), where \(j = 6\), \(n = 7\). The 2nd hypothesis is \(\{g(7)\}\), since \(j+i-1 \leq n\) (6+2-1 \(\leq 7\)). Finally the 3rd hypothesis is \(\{g(5), g(6), g(7)\}\) because \(j+i-1 > n\), and \(K = 6+3-1-7 = 1\).

\[
g(0) g(1) g(2) g(3) g(4) g(5) g(6) g(7)
\]

\[
g(2) g(10) g(14) g(26) g(46) g(73)
\]

\[
1 3 4 2
\]

Figure 8. Process of sequential clustering and the ranked list of hypotheses

IV. Experimental results

We have used a set of 498 binary images of handwritten Korean text lines to evaluate the performance of gap metrics and clustering techniques. The data are constructed by extracting the address lines from live mail pieces in Korea, and the locations of between-word gaps are labeled manually. Our objective is to obtain the best combination of gap metric and clustering technique for handwritten Korean word segmentation. The experiment has been run on a Pentium III 355MHz PC.

4.1 Evaluation of gap metrics

A quantitative measure to evaluate the performance of gap metrics has been proposed in [4, 5]. A gap metric is said to have ordered a line correctly, if the sorted list of gaps contains all word gaps after any non-word gap. Therefore the percentage of line images that are correctly ordered designates how good a gap metric is. Table 1 compares the performance of the gap metrics considered in Section 2 as well as their combination, \((\text{BB, RLE, CH})\). Here \((\text{BB, RLE, CH})\) means that a gap size is viewed as a three-dimensional vector whose components are BB, RLE, and CH metrics.

<table>
<thead>
<tr>
<th>Gap metric</th>
<th>% of correct ordering</th>
</tr>
</thead>
<tbody>
<tr>
<td>BB</td>
<td>93.98</td>
</tr>
<tr>
<td>RLE</td>
<td>92.77</td>
</tr>
<tr>
<td>CH</td>
<td>94.58</td>
</tr>
<tr>
<td>(BB, RLE, CH)</td>
<td>94.98</td>
</tr>
</tbody>
</table>

The Venn diagram in Figure 9 shows the distribution of incorrect orderings caused by BB, RLE, and CH for the data set. An incorrect ordering happens when a distance measure over-estimates within-word gaps and/or under-estimates between-word gaps. Although the individual metrics have an error rate of 5-8%, a relatively small percentage (19/498=3.8%) of the lines are incorrectly handled by all the three gap metrics. This suggests that an intelligent combination of the metrics can yield a better performance. (BB, RLE, CH) is one of the possible combinations and has a superior performance – see Tab. 1.

Figure 9. Distribution of incorrect-ordering

In [5], an evaluation of BB, RLE and CH using Roman-style text line images is conducted, and the performance was ordered as CH, RLE, and BB. Our experiment also produces a similar result. The BB and RLE measures are all simple and efficient, but they have some drawbacks. As can be seen from Figure 10, the BB metric can underestimate a between-word gap (gap ‘5’) while the RLE metric can over-estimate a within-word gap (gap ‘2’). In this case, only the CH measure generates a correct ordering.
4.2 Performance of word segmentation

To accomplish the task of word segmentation, a gap metric is used to measure the size of gaps and a clustering technique is applied to classify every gap into a BWG or a WWG. There are 12 possible combinations – four gap metrics and three clustering techniques. Figure 11 compares the performance of the 12 combinations. We denote the average linkage method by ALM, the modified MAX method by mMAX, the sequential clustering by SC, and (BB, RLE, CH) by ALL for convenience. We have obtained the best performance by applying the sequential clustering using all the three gap metrics, where the cumulative accuracies up to the 3rd hypothesis are 80.52%, 92.57%, and 94.98%, respectively. For this combination, the processing time is 0.5 second per image.

Figure 10. Comparison of the BB, RLE and CH measures

Figure 11. Accuracies up to the 3rd hypothesis of every combination

Note the fact that the cumulative accuracy of (ALL, SC) up to the 3rd hypothesis is 94.98% which is the percentage of correct ordering of the gap metric ALL. It means that once the gap metric produces a correct ordering, the clustering technique finds the correct segmentation in one of the three hypotheses.

V. Concluding Remark

We have proposed a gap clustering method that can divide a handwritten Korean text line image into words. We take up three kinds of gap metrics – BB, RLE and CH that have produced superior performance in Roman-style word segmentation, and three clustering techniques – average linkage method, modified MAX, and a sequential clustering to classify the gaps into between-word gap or within-word gap. An experiment with 498 text line images from live mail pieces shows that the combination of the sequential clustering technique, proposed in this paper, and a gap metric with all the three measures has the best performance – 80.52% of accuracy and 0.5 second of processing time per line.

ACKNOWLEDGEMENT

This work was supported by the Brain Korea 21 Project and grant number 98-0102-02-01-3 from the interdisciplinary Research program of the KOSEF.

REFERENCES