Word Segmentation of Printed Text Lines Based on Gap Clustering and Special Symbol Detection

Soo H. Kim[*], Chang B. Jeong[*], Hee K. Kwag[**], Ching Y. Suen[***]

[*]*Department of Computer Science, Chonnam National University, Korea*
shkim@chonnam.chonnam.ac.kr
[**]*Center for Artificial Intelligence Research, KAIST, Korea, hkkwag@ai.kaist.ac.kr*
[***]*Center for Pattern Recognition and Machine Intelligence, Concordia University*
Canada, suen@cenparmi.concordia.ca

Abstract

This paper proposes a word segmentation method for machine-printed text lines. It utilizes gaps and special symbols as delimiters between words. A gap clustering technique is used to identify the gaps between words regardless of the gap-size variations among different document images. Next a special symbol detection technique is applied to find two types of special symbols lying between words. An experiment with 1,675 text lines in 100 different English and Korean documents shows that the proposed method achieves a high accuracy of word segmentation.

1. Introduction

Segmentation of a printed text line into words is an essential component of a variety of document manipulation systems, such as optical character reader (OCR), document imaging system, keyword spotting system for digital library, and so on. However, only a few research results on this issue can be found in the literature [1, 2].

Existing methods for word segmentation adopt a gap-based approach which assumes that there is a significant gap between adjacent words. But, in practical situations, there are many difficulties. First, the magnitude of word gap varies considerably from one image to another depending on the scanning resolution, language in the document, word processor and font style, and so on – refer to Figure 2 where the size of a word gap is quite different between Korean and English text lines. This fact implies that accurate gap-based word segmentation should be adaptive to these variations. Second, the gap is not the only delimiter between words: special symbols such as dashes and parentheses can also be used to delimit the word – refer to Figure 4 in which words are delimited by special symbols instead of a significant gap.

In this paper we propose a word segmentation method composed of gap clustering and special symbol detection techniques. It utilizes gaps and special symbols as delimiters between words. This method was evaluated by 1,675 text lines of 100 different English and Korean documents, and an encouraging accuracy of 99.83% was produced.

2. Gap clustering

2.1. Gap metric

In Roman-style word segmentation, the gap is defined as a white space between two adjacent *connected components* (CC's), since the text line can be regarded as a one-dimensional stream of CC's running from left to right [3, 4, 5]. For Korean text lines, the gap should be defined differently because a character is a two-dimensional combination of consonants and vowels. Therefore we define the gap as a white space between two adjacent *overlapped components* (OC's), where an OC is defined as a set of CC's whose projection profiles in the vertical direction overlap. In this regard, a text line is a one-dimensional stream of OC's [6].

1051-4651/02 $17.00 (c) 2002 IEEE

Given an English or Korean text line, we detect the 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 a sample text line containing two types of gaps: between-word gap (BWG) and within-word gap (WWG).

Figure 1. Gaps from a vertical projection

A number of gap metrics to measure the size of gaps have been proposed for the segmentation of handwritten Roman-style words [3, 4, 5]. Considering the rectangular nature of Korean characters, we adopt a bounding box (BB) metric [3]. Actually all the gap metrics have been tried during the development of our method and the BB metric has produced the best results.

The BB measure computes the gap size as the horizontal distance between the bounding boxes of adjacent OC's. A bounding box of an OC is formed by grouping a number of CC's which overlap each other in the vertical direction. A CC whose size is smaller than a threshold is excluded from the grouping, because it is either noise or a punctuation mark such as comma, quotation mark, and so on. Figure 2 shows the gap distances computed by the BB measure.

2.2. Average linkage clustering

The problem of gap-based 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). For a set of gaps $\{g_1, g_2, \cdots, g_n\}$, the function f can take one of three forms: (1) it maps every g_i to the WWG if the line image contains only one word, (2) it maps every g_i to the BWG if the image contains n+1 words and all the characters within a word overlap, and (3) it maps some g_i's to the WWG and the others to the BWG.

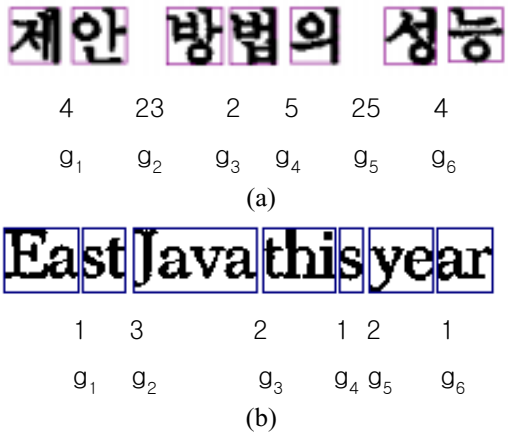


Figure 2. Gap distances computed by the BB metric: (a) Korean text line, (b) English text line

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 the 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 type from the second, the normalized mean is used. All the gaps are classified into the WWG if the mean is smaller than a threshold; otherwise all of them are classified into the BWG. The thresholds are determined empirically. For the third case, we consider the following clustering technique to determine the membership of each gap.

The average linkage method is one of hierarchical agglomerative clustering techniques [5]. 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,

$$D(C_i, C_j) = 1/(n_i n_j) \sum_{a \in Ci,\, b \in Cj} D_m(a, b),$$

where n_i (or n_j) is the number of gaps in a cluster C_i (or C_j), and D_m indicates a function of Manhattan distance.

The result is a dendrogram like the one shown in Figure 3 which illustrates the process of merging six clusters into one using the BB distances in Figure 2(a). It is organized in such a way that the right cluster of

any node has a larger mean than its left cluster. The root node (node 'e' in Figure 3) can be used to produce the most probable decision of word segmentation – all its in-order successors are BWG's, i.e., the two gaps whose sizes are 23 and 25 in Figure 3 are BWG's.

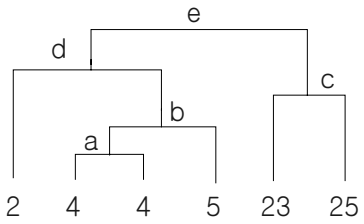


Figure 3. Dendrogram as a result of clustering with the gaps shown in Figure 2(a)

3. Special symbol detection

We detect two types of special symbols using a rule-based algorithm without involving the recognition of any symbols. The first one is for the special symbols which are elongated horizontally such as dash ('–') or tilde ('~'). The second one is for the vertically elongated symbols such as various kinds of parentheses '{', '}', '[', ']', '(', or ')'. These symbols can be used for further decomposition over the result of gap-based segmentation – see an example in Figure 4. Detection of the two types of special symbols is performed by a rule-based algorithm.

Figure 4. Special symbols for word segmentation

(Rule 1) For the i-th OC with width W_i and height H_i respectively, if it satisfies both the following two conditions, it is regarded as a special symbol of the first type:

(1) $W_i > 2 \times H_i$

(2) $H_i < T_1$, where T_1 is 30% of the median height among all OC's

(Rule 2) On the other hand, if the i-th OC satisfies the following four conditions, it is regarded as a special symbol of the second type:

(1) $H_i > 2 \times W_i$

(2) $W_i < T_2$, where T_2 is 30% of the median width among all OC's

(3) $D_i < 0.75 \times W_i \times H_i$, where D_i is the number of black pixels of the OC

(4) The upper 10% and lower 10% of the OC are symmetric AND the left and right parts are not symmetric

Conditions (1) and (2) in Rule 1, and conditions (1)-(3) in Rule 2 find special symbols using some syntactic features, while condition (4) in Rule 2 differentiates a character or a part of a character from the special symbol.

4. Experiments

We have implemented the proposed method in a C++ programming language on a P-III 450MHz PC, and evaluated its performance with 1,675 text lines contained in 100 text blocks from a variety of English and Korean documents scanned at 300 DPI – see Table 1.

Table 1. Test data for experiment

	Korean Document		English Document		Total
	Jour.	Mag.	Jour.	Mag.	
#-blocks	80	5	10	5	100
#-lines	1,244	98	151	182	1,675
#-words	7,401	474	1,532	1,140	10,547

Every text block is first divided into lines by a recursive X-Y cut method [7]. As a result, 1,673 out of 1,675 text lines are extracted successfully. Two adjacent text lines in a block of English magazine have been extracted as one line, but we divided it manually.

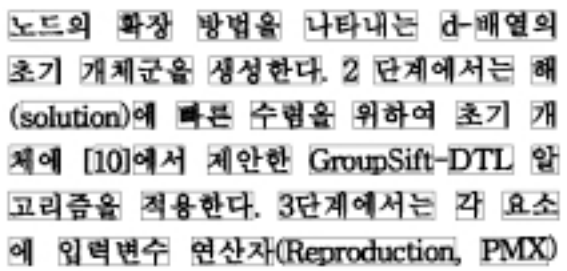
Given an image of text line, gap clustering (GC) is applied first. Measuring the accuracy of word segmentation as the ratio between the number of words segmented correctly and the total number of words, gap clustering as a word segmentation methodology produced an accuracy of 96.9%. After special symbol detection (SSD) on the result of gap clustering, the accuracy improved up to 99.83% – see Table 2. There are 18 errors in word segmentation, and most of the error dues to the mistake of SSD:

there are 639 special symbols in our test data, and the SSD makes 10 mistakes (missing 8 special symbols, and mis-classifying 2 characters as special symbols) and has an accuracy of 98.44%.

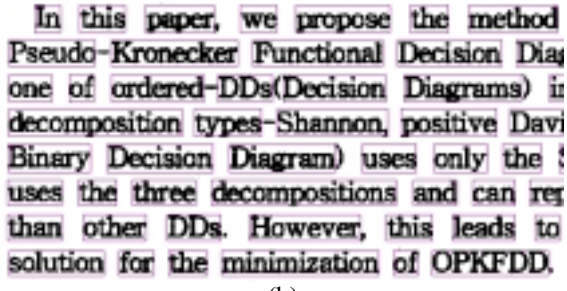
Figure 5 shows some examples of word segmentation by the proposed method. Note that the special symbols as well as gaps are used to delimit words. In addition, small components are ignored, as described in Section 2.1, because they are either noise or punctuation marks such as commas, periods, quotation marks, and so on.

Table 2. Accuracy of word segmentation (%)

	Korean Document		English Document		Total
	Jour.	Mag.	Jour.	Mag.	
#-word	7,401	474	1,532	1,140	10,547
after GC	96.33	98.31	98.30	98.15	96.90
after SSD	99.93	99.36	99.80	99.38	99.83



(a)



(b)

Figure 5. Example of word segmentation: (a) Korean document, (b) English document

5. Conclusion

A word segmentation method composed of gap clustering and special symbol detection techniques has been proposed. The gap clustering technique is immune to the variations of scanning resolutions, languages in the document, word processors and font styles, and so on. The special symbol detection technique finds two types of word-delimiting symbols using syntactic features without having to recognize any individual symbols. The proposed method has been applied to the segmentation of Korean and English documents and proven to be effective. We are extending the gap clustering technique to deal with multi-lingual text lines.

Acknowledgement

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

References

[1] Y. Y. Tang, S. W. Lee, and C. Y. Suen, "Automatic Document Processing: a Survey," *Pattern Recognition*, Vol. 29, No. 12, pp. 1931-1952, 1996.

[2] K. Chung and H. Kwon, "A Feature-Based Word Spotting for Content-Based Retrieval of Machine-Printed English Document Images," *Journal of Korea Information System Society*, Vol. 26, No. 10, pp. 1204-1218, 1999. (text in Korean)

[3] G. Seni and E. Cohen, "External Word Segmentation of Off-line Handwritten Text Lines," *Pattern Recognition*, Vol. 17, No. 1, pp. 41-52, 1994.

[4] U. Mahadevan and R. C. Nagabushnam, "Gap Metrics for Word Separation in Handwritten Lines," *Proc. 3rd Int. Conf. Document Analysis and Recognition*, pp. 124-127, Montreal, Canada, 1995.

[5] U. Mahadevan and S. N. Srihari, "Hypotheses Generation for Word-Separation in Handwritten Lines," *Proc. 5th International Workshop on Frontiers in Handwriting Recognition*, pp. 53-456, Essex, England, 1996.

[6] S. H. Kim, S. Jeong, G. S. Lee, and C. Y. Suen, "Word Segmentation in Handwritten Korean Text Lines Based on Gap Clustering Techniques," *Proc. 6th Int. Conf. Document Analysis and Recognition*, pp. 189-193, Seattle, USA, 2001.

[7] J. Ha, R. M. Haralick, and I. T. Phillips, "Recursive X-Y Cut using Bounding Boxes of Connected Components," *Proc. 3rd Int. Conf. Document Analysis and Recognition*, pp. 952-955, Montreal, Canada, 1995.