Reconstruction of Overlapping Character in Thai Printed Documents

Nucharee Premchaiswadi*, Wichian Premchaiswadi*,
Voravit Premratanachai*, and Seinosuke Narita**

* Faculty of Information Technology,King Mongkut's Institue of Technology Ladkrabang,
Bangkok, 10520, Thailand
Tel. (66) 2-7372551-4 Ext.540 Fax. (66) 2-3269047
E-mail: nucharee@it.kmitl.ac.th
** Department of Electrical, Electronics and Computer Engineering,
School of Science and Engineering, Waseda University, Tokyo, Japan
Tel. (81) 3-5286-3374 Fax. (81) 3-3202-9940
E-mail: narita@narita.elec.waseda.ac.jp

Abstract: This paper proposes a reconstruction scheme for overlapping characters in Thai printed document. Overlapping characters are characters that overlap with surrounding characters. The problem of overlapping characters is still an unsolved problem in commercially available software of Thai character recognition systems. The algorithm of reconstruction scheme is based on structural analysis of overlapping Thai printed characters. It consists of 2 steps: overlapping point determination and reconstruction of segmented characters. The overlapping point is defined as the intersection point between characters and can be determined by using templates. Then, an overlapping character is separated into segments at the intersection point. The structure of each segment may be an incomplete character and is not identical to the original one. Therefore, the reconstruction process is employed to add the incomplete part of these segments. The proposed scheme has been implemented and tested with 70 patterns of conventionally found in overlapping printed Thai characters with different typefaces and type sizes. The experimental results show that the proposed scheme can segment and reconstruct overlapping characters correctly. The proposed scheme can improve the recognition rate of commercially available software, ThaiOCR1.5 and ArnThai1.0, more than 60 percents.

1. Introduction

There is the case that one character will overlap with another character. Overlapping characters usually occur from type font and capability of word processors in allowing proportional display or printing. Although human can read and understand these characters. The recognition process can not recognize these characters correctly because two characters are combined and considered as a single character. The problem of overlapping Thai character is one of the most important problems that decrease recognition rate in a character recognition system. It is still an unsolved problem in commercially available software of Thai character recognition systems. The problem is different from the problem of touching characters. The touching character problem concerns with the connection between characters while the overlapping characters problem concerns with the overlap between characters. Therefore, the algorithms for touching character segmentation [1]-[3] can not be applied to overlapping characters properly. In case of overlapping characters, the segmentation scheme for touching characters will produce the incomplete segmented characters that are difficult to recognize or can not be recognized at all. One of the segmented characters may miss some part of the character while the other may have part of other character connected with. To solve the problem, this paper presents characteristics of overlapping Thai characters and a scheme for segmentation and reconstruction for the overlapping characters.

2. Characteristics of Overlapping Characters

The multi level structure of Thai sentence composed of three types of characters namely: consonants, vowels and voice tones. However, these characters may locate in different level. According to the location of these characters, they can be divided into three zones namely: the upper zone (UZ), central zone (CZ) and lower zone (LZ) [3], as shown in Figure 1.

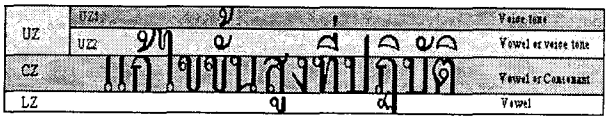

Figure 1. Zone classification and type of Thai characters.

The investigation on possible overlapping characters in daily used type fonts, overlapping characters permitted in Microsoft Word Processor and overlapping characters available in Thai language found that overlapping characters appeared in 70 patterns, as shown in Table 1. And all of these patterns occurred only in the upper zone. Therefore, the algorithm for overlapping characters needs to consider only in the upper zone. From the 70 patterns, we can group them into two types as samples shown in Figure 2 as follows:

1) overlapping of consonant characters with vowel characters and
2) overlapping of vowel characters with vowel characters

Table 1. Overlapping characters in 70 patterns.

ปัฝฟ	ริ	รี	รึ	รื	ริ	รี
ปัฝฟ	ริ	รี	รึ	รื	ริ	รี
ปัฝฟ	ริ	รี	รึ	รื	ริ	รี
พิ	ริ	รี	รึ	รื	ริ	รี
ริ	ริ	รี	รึ	รื	ริ	รี
ริ	ริ	รี	รึ	รื	ริ	รี
ริ	ริ	รี	รึ	รื	ริ	รี
ริ	ริ	รี	รึ	รื	ริ	รี
ริ	ริ	รี	รึ	รื	ริ	รี
ริ	ริ	รี	รึ	รื	ริ	รี

*Symbol "ร" is a consonant character that is needed to display voice tone and vowel character by Thai word processing.

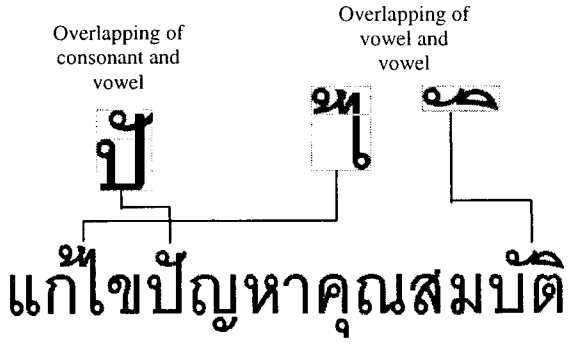

Figure 2. An example of overlapping characters.

3. Boundary of Characters an Overlapping Types Classification

The boundary of each zone can be found by using horizontal projection [3] as shown in Figure 3. The height of character (HC) in each line is also obtained.

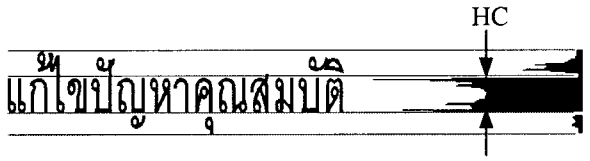

Figure 3. Horizontal projection.

Then, the boundary of each connected pixel called character frame is obtained by using an edge detection algorithm [4]. A character frame may consist of a single character or touching characters or overlapping characters. In the paper, only the overlapping character is considered. The solution of touching characters can be found in other researches [1]-[3].

The two types of overlapping characters can be classified by using distinctive features of characters. If the character frame occupy more than one zone, it will be separated into two character frames: one frame in upper zone and another in central zone by using the height of that line. Then, the first type of overlapping character can be detected by checking the width of character frame in the upper zone and the central zone. The following two conditions are used to classify this type of overlapping characters.

1) The width of character frame in the central zone is greater than 0.8*(AW), where AW is the average width of the character which is equal to 0.8*(HC).
2) The width of the upper zone is greater than 0.5*(AW).

In case of the second type, it can be subdivided into 2 cases as shown in Figure 4. The first case can be detected by checking the width in the upper zone and it must greater than 1.45*(AW). The second case is detected by checking the width of character frame in the upper zone and central zone. The width of character frame in upper zone must greater than AW and the width of central zone must less than 0.8*(AW).

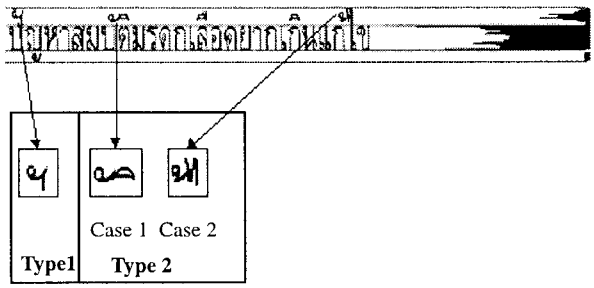

Figure 4. Type of overlapping characters.

4. The Proposed Reconstruction Scheme

The proposed reconstruction scheme consists of two steps: overlapping points determination and reconstruction of segmented characters as follows.

4.1 Overlapping points determination

The overlapping point is defined as the intersection point between characters and will be used for separating overlapped parts of characters. The overlapping point is determined by using templates, which are patterns of pixel [4]. According to the types of overlapping characters, the templates are also divided into two types. For finding the overlapping point of the first type, four templates as shown in Figure 5 are employed. The characteristics of overlapping character of type 1 are varies depended on type fonts of characters. There are two forms of overlapping character appearance. Form one, tail of a character appears as shown in Figure 6(a) and 6(b). Form two, tail of a character disappears as shown in Figure 6(c).

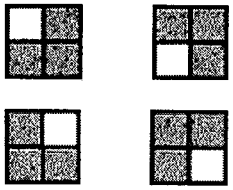

Figure 5. Four templates used for type 1.

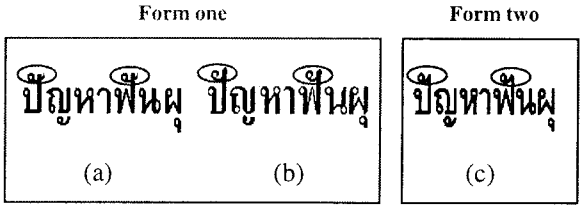

Figure 6. Form of overlapping characters appearance in type 1.

For explanation purpose, character leg is a vertical line appeared on a character. The selected character leg is determined by two maximum vertical projection of character frame in upper zone that corresponds with central zone. The overlapping point is found by scanning vertically along the left and right edge of selected character leg (Sx and Ex in Figure 7). The four templates are scanned from top to bottom in vertical direction as shown in Figure 8.

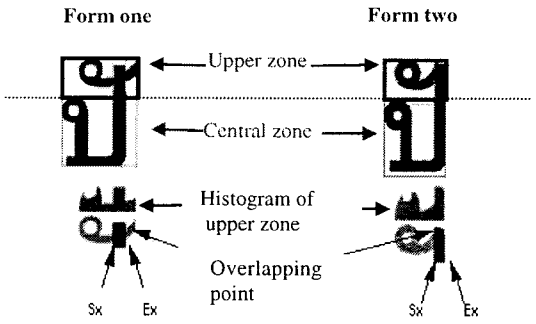

Figure 7. Finding edges of character leg.

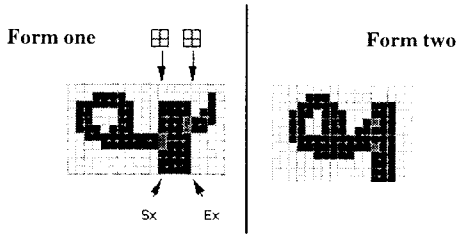

Figure 8. Finding overlapping point.

In the second type, only one template is used as shown in Figure 9. P0, P6 and P7 are black pixels. Q and P2 are white pixels. P1, P3, P4 and P5 are ignored pixel.

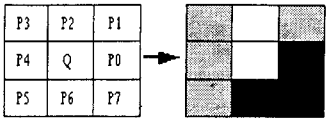

Figure 9. The template used for type 2.

The boundary for scanning in vertical direction is obtained from the two highest peaks of horizontal projection of the character frame as shown in Figure 10.

Figure 10. Boundary for vertical scanning.

The starting and ending point of the scan in horizontal direction is determined as follows:

Start = (hisY_max – hisY_min)
End = 1.2 * (his_e – his_s)/2.

Figure 11 and Figure 12 show an example of finding a staring and ending point, respectively. The overlapping point is found by horizontally scan from right to left boundary. The scanning starts from the minimum point of horizontal projection as shown in Figure 11. If the pattern of the template can not be found, the position of scan will be increased by one pixel as show in Figure 13.

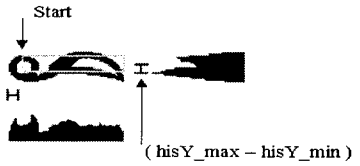

Figure 11. Finding the starting point.

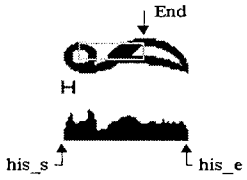

Figure 12. Finding the ending point.

Figure 13. Change level for finding crossing point

The segmentation process uses the overlapping point for separating a overlapping character into two segments. The structure of each segmented character may be an incomplete character and may not identical to the original one.

4.2 Reconstruction of segmented characters

The proposed reconstruction process is used to add the incomplete part of these segmented characters. The reconstruction can be categorized into two types according to the types of overlapping character as shown in Figure 4. The reconstruction scheme for the first type is divided into two cases according to two forms of overlapping appearance as shown in Figure 14. In the first case, four overlapping points will be found while in the second case has only two overlapping points.

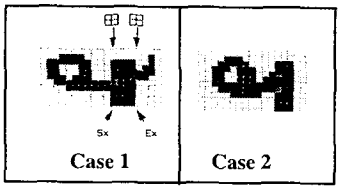

Figure 14. Two cases of type 1.

In the first case, the reconstruction procedure is performed by connect the pair of crossing points and delete the pixel which is outside the boundary as shown in Figure 15.

Figure 15. Reconstruction of type 1.

The reconstruction of the second case has the same operation as that of the second type. In this case, we need to add the tail of character from the two crossing points obtained from the previous process. The incomplete part must be constructed by using only the information of the structure of the character itself. The important factors of the reconstruction process are slope and length of the incomplete part. The slope is calculated from two points. The first point is the upper point of the cutting point, and the second point is the left point of the longest horizontal part in segmented character as show as in Figure 16 (a).

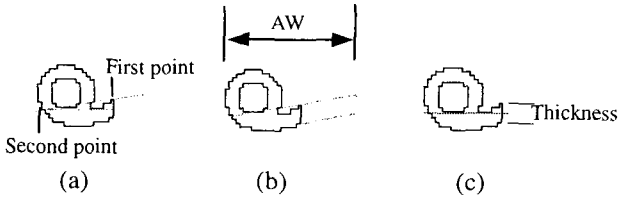

Figure 16. Reconstruction of the second type.

The length of the incomplete part is calculated from the different between AW and the width of segmented characters as shown in Figure 16 (b).

The incomplete part usually be a tail of character. Therefore, the thickness of the tail can be calculated from the thickness of the character at the cutting point as shown in Figure 16 (c).

5. Results and Conclusion

The proposed scheme is implemented by using Visual C++ version 6.0. The experimentation of this scheme was performed with 70 patterns of overlapping characters and some examples are shown in Figure 17.

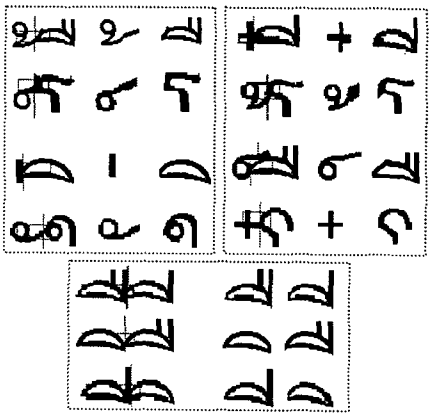

Figure 17. An example of experimental result.

The results show that the proposed scheme can segment overlapping characters correctly. The proposed scheme can improve of correctness of commercially available software, ThaiOCR1.5 and ArnThai1.0, more than 60 percents as show in Table 2.

Table 2. Percentage of correct recognition.

Software	Before	After
ArnThai 1.0	10 %	83.33%
ThaiOCR 1.5	20.83 %	82.5 %

References

[1]Yi Lu, "Machine Printed Character Segmentation-An Overview", Pattern Recognition, Vol.28, No.1, pp. 67-80, 1995.

[2]Richard G. Casey and Eric Lecolinet, "A Survey of Methods and Strategies in Character Segmentation", Proceeding of IEEE Vol.18, No.7, pp.690-695, July 1996.

[3]Nucharee Premchaiswadi, Wichian Premchaiswadi, Seinosuke Narita, "Segmentation Of Horizontal and Vertical Touching Thai Character", ITC-CSCC'99 International Technical Conference on Circuit Systems, Computer and Communications, Niigata, Japan, 1999.

[4]Rafael C. Gonzalez and Richaed E. Wood, Digital Image Processing, Addison-Wesley, 1999.