Optimization of Line Segmentation Techniques for Thai Handwritten Documents

Olarik Surinta

Abstract— **The purpose of the research is to study the optimization of line segmentation techniques for Thai handwritten documents. This research considered only single-column of Thai documents. I proposed two new techniques including comparing Thai character and sorting and distinguishing. These two techniques were used with recognized techniques on the basis of projection profile (including horizontal projection profile and stripe) in the experiment. The outcome of this research suggested that the best technique for single-column Thai documents is the new technique for sorting and distinguishing, this technique provide the accuracy of 97.11%**

I. Introduction

In handwritten recognition, the line segmentation is an essential scheme. This is because the occurrence of an inaccurately line segmentation will cause errors in the character segmentation. Several techniques of line segmentation have been proposed in the past, most of line segmentation techniques have been based on horizontal projection profile technique. This is because of the texts in most document images are aligned along horizontal lines.[1] The projection profile technique is mostly used for segmenting characters, words, and text lines. [2]

Projection profile based techniques [2, 3] may be one of the most successful top-down algorithms for printed character documents [4] since the gap between two neighboring text lines in printed character documents is typically significant, thus the text lines are easily separable. However, these projection profile techniques cannot be directly used in handwritten documents, unless gaps between the lines are significant or the handwritten lines are straight. [5]

This paper is organized as follows. Section 2 describes the nature of Thai language. In Section 3, I describe how the horizontal projection profile to segments Thai handwritten documents into character lines. Section 4 presents the stripe technique for horizontal projection profile. Section 5 presents a technique used for comparing between consonant and a group of small vowel and tone based on horizontal projection profile. Section 6 presents a new technique for sorting and distinguishing on the basis of projection profile. Concluding remarks are given in Section 7.

Manuscript received May 14, 2009.

O. Surinta is with the Department of Management Information Systems, Faculty of Informatics, Mahasarakham University, Mahasarakham, 44150 Thailand (phone: +66-43-754-359; fax: +66-43-754-359; e-mail: olarik.s@msu.ac.th).

978-1-4244-4139-6/09/$25.00 ©2009 IEEE

II. The Characteristic of Thai Character

Thai characters consist of 46 consonants, 18 vowels and 4 tones, [6, 7] as shown in Table 1. Thai characters from Table 1 are used together to make the sentences in Thai Language. Thai writing style is from left to right [8]. Thai character utilizes a concept of upper or lower characters. As a result, the Thai sentences can be divided into 4 levels [6, 9] such as lower level, centre level, upper level 1 and upper level 2, as shown in Fig. 1.

According to the characteristic of Thai character, consonants are generally presented in the centre level; however some consonants such as “ฟ” ,“ฝ” and “พ” can be present in the centre level and upper level 1. In addition, vowels can be presented in the lower level, centre level, upper level 1 and upper level 2. Finally, Tones can be presented in the upper level 1 and upper level 2.

The size of tone and vowel in the upper level and lower level of the Thai sentence structure are generally smaller than consonants around 30 percents. Consequently, I proposed two new techniques of line segmentation from Thai handwritten image document including technique for comparing Thai character, as shown in Section 5 and technique for sorting and distinguishing, as shown in Section 6. These techniques based on the basis of horizontal projection profile.

TABLE I
Thai Characters

Character Types	Character
Consonants	ก ข ฃ ค ฅ ฆ ง จ ฉ ช ซ ฌ
	ญ ฎ ฏ ฐ ฑ ฒ ณ ด ต ถ ท ธ
	น บ ป ผ ฝ พ ฟ ภ ม ย ร ฤ
	ล ฦ ว ศ ษ ส ห ฬ อ ฮ
Vowels	อั อะ อา อิ อี อึ อื อุ อู
	เอ โอ ใอ ไอ ๅ ๆ อี๋ อ์ อำ
Tones	อ่ อ้ อ๊ อ๋

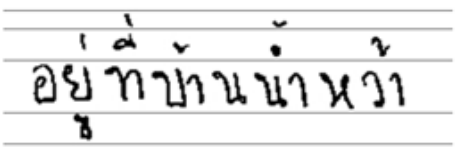

Fig. 1. Thai sentence structure.

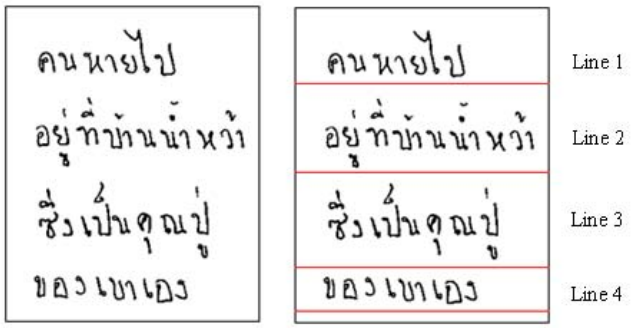

Fig. 2. Line segmentation Thai handwritten document.

III. The Horizontal Projection Profile to Segment Thai Handwritten Documents into Character Lines

The horizontal projection profile, which is a basic of line segmentation technique, [3, 5, 10] is used in dividing the text image into character line. In our approach the horizontal projection profile is obtained by summing pixel values along the horizontal axis (x-axis) for each row (y-axis) row. The vertical gaps between the text lines can be determined, as shown in Fig. 3.

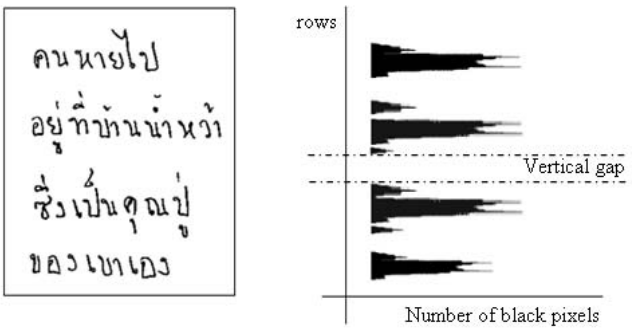

Fig. 3. Horizontal projection profile of Thai handwritten document.

Horizontal projection profile uses the vertical gaps between black pixels to separate the character line from document images. After applying this technique to our example in Fig.3, 7 character lines are obtained, as shown in Fig. 4. Actually, the image of this document contains 4 lines of sentences, as shown in Fig. 2.

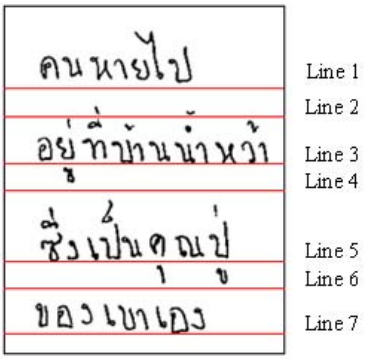

Fig. 4. The result of horizontal projection profile.

IV. The Stripe Technique for Horizontal Projection Profile

The stripe technique presented by Tripathy, N. and Pal, U. [11]. They proposed for Oriya text image document. This research provides the line segmentation of Oriya text accuracy to 97%. This is because of vowels and consonants are written in the same level.

I apply the strip technique to line segmentation of the Thai character image document. Firstly, the stripe technique divides image into stripe (small column) [11]. After that, the horizontal projection profile is used to divide the text image into character lines as mentioned in Section 3. The result of this technique is divides the image document into 9 lines, as shown in Fig. 5. This is because of this technique is suitable for skew documents of English characters.

The stripe technique is not coping well with Thai handwritten image documents. This is because of the horizontal projection profile cannot completely separate the character lines. It causes mistakes in the process of connecting the line marker between stripes.

Therefore I suggest our technique in order to cope with Thai handwritten image documents. This technique is comparing Thai characters between consonants and a group of small vowels and tones as mention in Section 5.

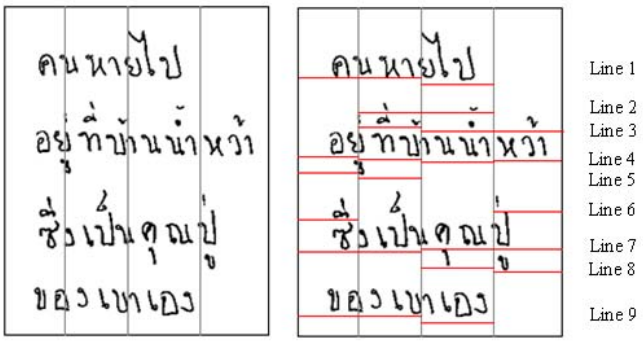

Fig. 5. The result of strip technique for horizontal projection profile.

V. Our Proposed Technique for Comparing Thai Character Based on Horizontal Projection Profile

This technique takes advantage of the differences in size of characters to differentiate Thai characters between consonants and a group of small vowels and tones. This is

because of the size of vowel such as “ิ”, “ี”, “ุ” and tone such as “่”, “้”, “๊” are smaller than consonants about 30 to 40 percents, as shown in Fig. 6.

Fig. 6. Comparing between consonant and a group of small vowel and tone.

There are two steps in this technique. First step is to study the groups of black pixels in the processed images. The groups are divided into two groups (upper and lower zone). The higher group (Fig. 7(a)) is then used to define the line from the image document. As a result, 4 line markers from first step are obtained. But, it cannot be completely separate the line segmentation from image document completely, for example, the line marker 2 and 3 from Fig. 7(b). This is because of vowels such as “ุ” and “ู” that are separated from consonant line.

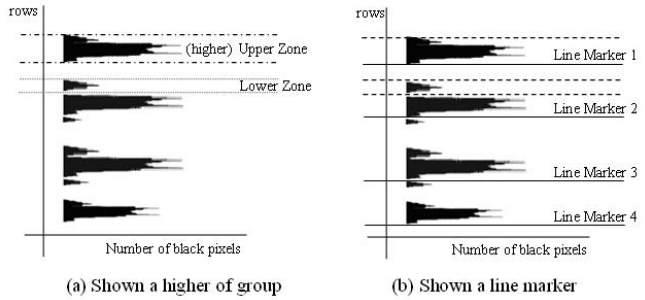

Fig. 7. The result of first step of comparison Thai character technique.

Second step, is to consider the high value of white pixel between the line markers and choose a new line marker, Figure 8 shows the result of using a new line marker. The process of choosing a new line maker can be time consuming and complicated. Thus, this technique is complex as there are many steps to be proved.

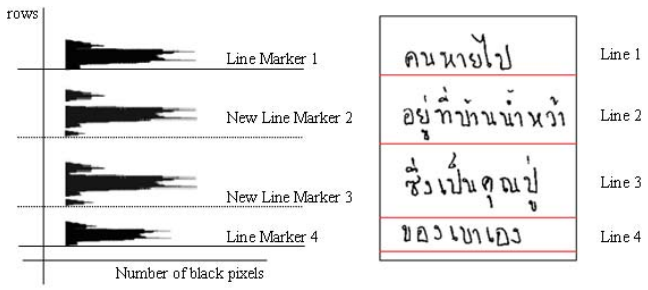

Fig. 8. The result of comparing Thai character technique.

VI. THE NEW TECHNIQUE FOR SORTING AND DISTINGUISHING ON THE BASIS OF PROJECTION PROFILE

In order to decrease the complexity in comparing the Thai character process mentioned in the section 5, the sorting and distinguishing on the basis of projection profile is proposed as a new technique to segment lines of sentences from image document. Moreover, this technique is not complicated and suitable for Thai character.

Firstly, I use the histogram of horizontal projection profile to sort the group of black pixels by starting with the minimum to maximum of number of black pixel in the right order, as demonstrated in Fig. 9.

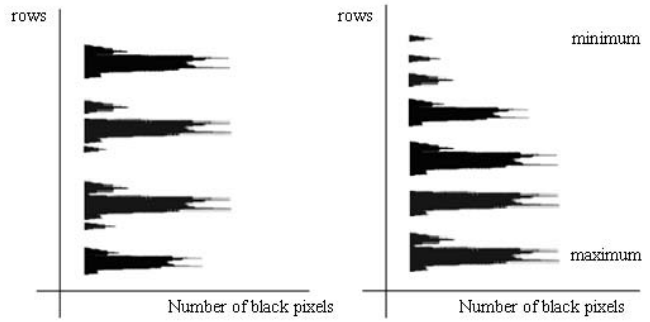

Fig. 9. Sorting the group of black pixels.

Secondly, the maximum difference between two groups of black pixels is found. After that, according to the value of the group of black pixels and maximum difference value, the line marker is marked on the middle of the group of black pixels when the maximum difference value is less than value of the group of black pixels, as shown in Fig. 10(b).

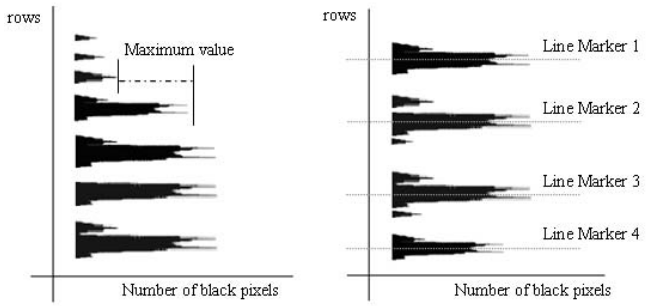

Fig. 10. The result of second step of sorting and distinguishing technique.

Finally, a new line marker is placed in the middle between every two conjunction line markers. The new line markers complete the process of separating lines in the paragraph from the image document as shown in Fig. 11.

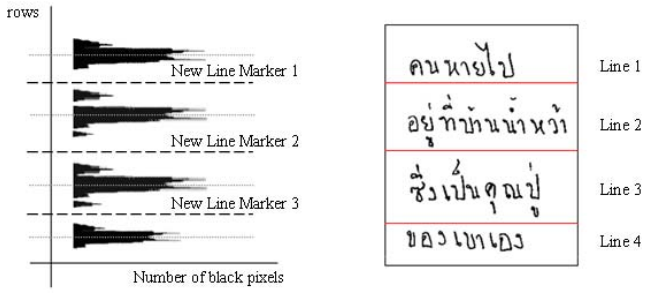

Fig. 11. The result of sorting and distinguishing technique.

VII. Experimental result

This section present the experimental result based on four line segmentation techniques, including horizontal projection profile technique, stripe technique, comparing Thai character technique, and sorting and distinguishing technique.

Thai image documents were generated from different peoples for the experiments of line segmentation technique. Therefore data sets contained varieties of writing styles, and limited to only single-column Thai image documents.

I manually calculated the accuracy of line segmentation. On the condition that, the line marker is used to define the character line; then, the line markers pass through the image document and do not cross the group of black pixels, consequently, line segment is completed. The results of line segmentation are shown in Table II.

TABLE II
THE RESULTS OF LINE SEGMENTATION

Number of lines on image documents	percentage			
	T1	T2	T3	T4
4	46	35	92	100
5	32	26	94	99
6	26	15	88	100
7	31	24	91	96
8	21	32	90	97
9	15	11	85	95
10	18	9	88	97
11	23	12	90	94
12	7	11	88	96
Average	24.33	19.44	89.55	97.11

T1 is Horizontal projection technique
T2 is Stripe technique
T3 is Comparing Thai character
T4 is Sorting and distinguishing

VIII. Conclusion

I have presented four techniques for the line segmentation of Thai language such as horizontal projection profile, stripe, comparing Thai character, and sorting and distinguishing. These 4 techniques based on horizontal projection profile. As a result, the horizontal projection technique provides the accuracy of 24.33%, the result of the stripe technique is less than the horizontal projection technique, the accuracy of this technique is 19.44%. Consequently, the horizontal projection profile technique and the stripe technique are suitable for English character and Oriya text.

In order to increase the accuracy, I proposed the comparing Thai character technique. The result of this technique increased the accuracy 65.25% from the horizontal technique. However, this technique is complex as there are many steps to be proved. Thus, I developed the new technique for line segmentation, it provides the Thai line segmentation accuracy of 97.11%. This technique is the sorting and distinguishing on the basis of projection profile.

References

[1] S. Li, Q. Shen, and J. Sum, "Skew detection using wavelet decomposition and projection profile analysis," *Pattern recognition letters*, vol. 28, pp. 555-562, 2007.

[2] O. D. Trier, A. K. Jain, and T. Taxt, "Feature extraction methods for character recognition - a survey," *Pattern recognition*, vol. 29, pp. 641-662, 1996.

[3] M. Zand, A. N. Nilchi, and S. A. Monadjemi, "Recognition-based segmentation in Persian character recognition," *International journal of computer and information science and engineering*, pp. 14-18, 2008.

[4] L. Likforman-Sulem, A. Zahour, and B. Taconet, "Text line segmentation of historical documents: a survey," *International journal on document analysis and recognition*, 2006.

[5] Z. Razak, K. Zulkiflee, M. Y. I. Idris, E. M. Tamil, M. Noorzaily, M. Noor, R. Salleh, M. Yaakob, Z. M. Yusof, and M. Yaacob, "Off-Line Handwriting Text Line Segmentation : A Review," *International journal of computer science and network security*, vol. 8, pp. 12-20, 2008.

[6] I. Methasate and S. Sae-tang, "The clustering technique for Thai handwritten recognition," presented at the 9th Int'l workshop on frontiers in handwriting recognition, 2004.

[7] N. Premchaiswadi, W. Premchaiswadi, U. Pachiyanukul, and S. Narita, "Broken characters identification for Thai character recognition systems," presented at Proceedings of the world scientific and engineering academy and society, 2003.

[8] A. Pornchaikajornsak and A. Thammano, "Handwritten Thai character recognition using fuzzy membership function and fuzzy artmap," presented at 2003 IEEE international symposium on computational intelligence in robotics and automation, Kobe, Japan, 2003.

[9] C. Tanprasert, W. Sinthupinyo, P. Dubey, and T. Tanprasert, "Improved Mixed Thai & English OCR using Two-step neural net classification," *NECTEC technical journal*, vol. 1, pp. 41-46, 1999.

[10] S. M. M. Mahmud, N. Shahrier, and A. S. M. D. Hossain, "An efficient segmentation scheme for the recognition of printed Bangla characters," presented at 7th International Conference on Computer and Information Technology Dhaka, Bangladesh 2004.

[11] N. Tripathy and U. Pal, "Handwriting segmentation of unconstrained Oriya text," vol. 31, pp. 755-769, 2006.