USING LOCAL MAXIMA PROFILE AND PIECE-WISE TECHNIQUE FOR LINE SEGMENTATION ON THAI HANDWRITTEN HISTORICAL DOCUMENTS

SEKSAN SANGSAWAD[1], RAPEEPORN CHAMCHONG[2], CHUN CHE FUNG[3]

[1, 2, 3] School of Information Technology, Murdoch University, Perth, Australia
E-MAIL: [1]seksan.s@gmail.com, [2]r.chamchong@ieee.org, [3]l.fung@murdoch.edu.au

Abstract:

This paper presents a new approach for segmenting text lines on Thai handwritten documents. The proposed technique is based on an Adaptive Local Connectivity Map concept using Piece-Wise Separating Lines. The algorithm is designed to solve problems in handwritten documents such as fluctuating text lines. Moreover, local maxima projection profile is used for enhancing the speed of extraction. The proposed algorithm consists of four steps. Firstly, Otsu algorithm is used to binarize the source image. Second, Piece-Wise Separating Lines is applied to derive the Adaptive Local Connectivity Map to show mask text lines. In the third step, local maxima projection profile is used as a guideline for extracting text lines. Finally, contour algorithm is used to identify the interested mask text line. The interested mask text is used to map with text image in order to extract the text lines. Analysis of experimental results on the King Rama 5 archive data indicated that the method has achieved a correct rate of 85.7%.

Keywords:

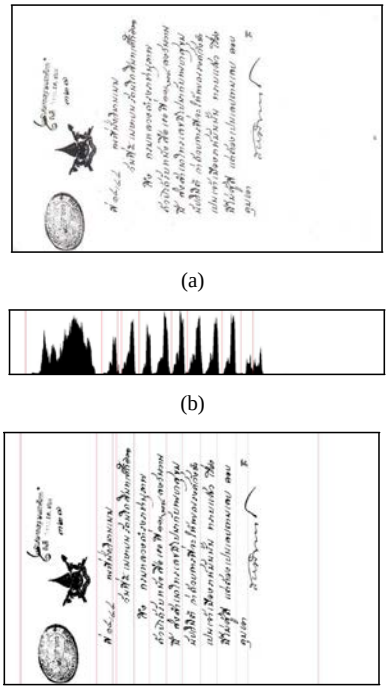
Thai handwritten document; Text line extraction; Local maxima; Piece-Wise Separating Lines (PSL); Adaptive Local Connectivity Map (ALCM)

1. Introduction

Many museums, libraries, and archives contain vast collections of handwritten historical documents. For example, an important collection is the archives from King Rama 5 preserved at the Nation Archive of Thailand. The archive of King Rama 5 of the 18th century has a significant value for Thai history. In the period of King Rama 5, Thailand experienced major changes in many aspects such as politics, governance and education. All the information had been well documented and archived. Most of valuable historical information was written within free-form handwritten documents. For retrieving the information, these documents need to be indexed in order to enable them to be searched efficiently.

Text line extraction is a necessary step for handwritten document recognition systems. The aim of the process is to segment an image into smaller text lines. In order to detect the lines, the areas of handwritten text are required to be recognized first in this step. Unlike fill-form documents such as cheques, postal or application forms, they have specific layout to facilitate the determination of the locations of handwritten text in the document. However, most sentences in historical documents were written with free-form handwriting which make localization and separation of the lines difficult. In addition, the handwritings are not always parallel to each other, implying that the spacing between the lines varies in the same page. Moreover, some lines of text are slightly curved. In addition, overlapping may occur due to ascender or descender characters. In addition, there are spaces separating the words in an English sentence while Thai words in a sentence are connected. There are also issues associated with multiple word levels for Thai writings. This causes much difficulty and challenges to perform segmentation and extraction for Thai language [11].

Normally, algorithms for text line separation emphasize on finding the location of the lines in order to segment them in their original logical orders. Many methods have been proposed in the literature. A well known technique for determining the direction of text lines is the Projection Profile technique [1-3] which creates a histogram crossing an entire text. In [6], the process of locating lines of text is focused only on the gap between the lines. Projection profile with local minima is used to detect the separating lines as shown in Figure 1. However, it could lead to a number of false local maxima and minima. Manmatha, et al. [7] addressed this problem by using the smoothed projection profile to reduce the sensitivity due to noise. In [4], projection profiles were used to determine the boundary between the baseline and then apply a contour following algorithm in that boundary area. Moreover, projection profile is used to determine the number of lines and estimate the beginning of the lines [5]. The corresponding space between the lines of the image is then examined to trace the boundary between the text lines. The trace of the line boundary may move up or down depending on the ascenders or descenders characters.



(a) (b) (c)

Figure 1. Steps of Line segmentation (a) Origianl image. (b) Projection profile. (c) Line segmentation by using local minima

Roy, et al. [8] used a piece-wise projection method to segment line in the document. The text is divided into vertical stripes of width and the width is computed from the average width of the characters within the dataset. Each stripe is computed separately in order to find the base of each line. The base line of the row is found where the sum of all the black pixels is zero. Although, the result of Piece-Wise Separating Lines (PSL) is not useful for line segmentation because lines in each strips are not connected, the distances between two consecutive PSLs in a stripe can be moved by using a statistical approach in order to get the individual boundary of each text line. Zhixin Shi, et al. [9] used an Adaptive Local Connectivity Map (ALCM) to process gray scale document images. First, the value at each pixel is computed by summing the value of the pixels in the gray scale image within a horizontal distance of that pixel. Next, binarization is applied in order to get the text line patterns with connected components. Then, grouping algorithm is used to define the location masks for each text line. Finally, the location mask block is mapped with the binary image for extracting text line.

In this paper, a new text line location and extraction algorithm for historical documents is proposed. The algorithm is designed for handling binary images. This proposed method applies piece-wise projection technique to perform as an adaptive local connectivity map in order to define the location masks of text line. In this technique, row-wise which defines the width of a row, will be placed when black pixel is found in each strip. After the process, the connectivity map will be generated. Moreover, local maxima projection profile is used for marking the center of each line. Local maxima projection profile speeds up the process in identifying the grouping component and location masks of text lines during individual line extraction.

This paper is organized as follows. Section 2 describes the steps in locating text lines and line extraction in detail. Section 3 presents the experimental results and Section 4 provides the conclusions.

2. Proposed Methodology

Two line detection techniques for locating and extracting text lines have been combined. The method consists the following steps.

(1) Standard thresholding algorithm on a gray scale document image to a binary image is applied.
(2) The piece-wise projection method is then perform to generate an adaptive local connectivity map for revealing the text lines.
(3) Local maxima projection profile is used for marking a center of each line in order to use it as the guideline for grouping connectivity components.
(4) Finally, the text lines from the document image can be extracted by mapping the location masks.

2.1. Binarization

Otsu [10] used global thresholding algorithm to transform gray scale document images to binary images. In [9], it has been shown that the algorithm works well for most historical manuscript images from the Library of Congress. It was also found that the backgrounds from the scan images of King Rama 5 archive are quite clear and low noise. This algorithm was therefore chosen. An example of before and after the processing stages is shown in Figure 2.

2.2. Location mask with PSL

Piece-wise projection was applied to detect the area of text line from sample documents among the King Rama 5 archive. In this method, the text is divided into vertical stripes of width. Width is the parameter for setting the size of the stripe and is calculated from the average width of the characters. 100 documents were chosen randomly and the

width was computed. Then, 10 random characters from each document were selected in order to calculate the average width. The width value of these dataset is 60 pixels. In the original PSL computation, the row-wise is the sum of all black pixels in horizontal direction of a stripe that equal to zero. However, in this proposed method the row-wise is placed when the sum of all black pixels is not zero. In this technique, the component from previous strip will be connected with component in current strip if they are in the same line. An illustration of connecting the component from two strips is shown in Figure 3. The location masks of text line will be appeared after this process is finished and it is shown in Figure 4.

Figure 2. The image on the left is the scanned image and the one on the right is the processed binary image from Otsu's Algorithm

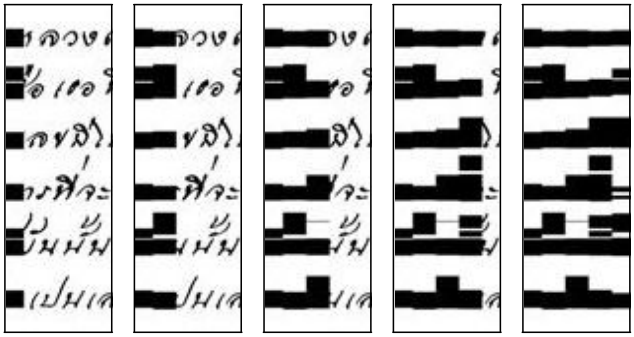


Figure 3. Sequence of connecting of component by applying PSL

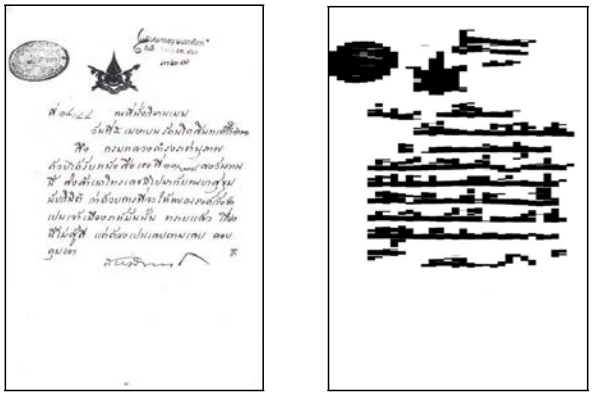


Figure 4. Image on the Left is the source image and right hand side one is a location mask image by applying PSL

2.3. Line detection with Local Maxima

Senior and Robinson [6] focused on the gaps between lines to determine the separating lines by using local minima. In this proposed method, we intend to find location of text line so that projection profile with local maxima is used to predetermine the center of each text line. The predetermined lines will pass through the location mask from the previous step. An example of predetermined lines with local maxima is shown in Figure 5. For locating text mask area, the mask document is scanned and prepared to be processed by the contour algorithm. The predetermined lines are used as the guidelines to point to the mask text lines by collecting the starting point of each line as an index. These guidelines will speed up the scanning process by pointing to the exact pixel line where the text mask text lines are located.

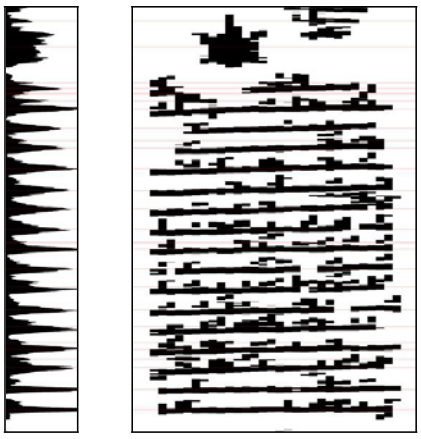
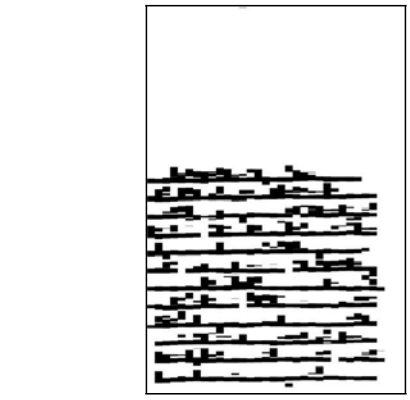


Figure 5. Left hand side is projection profile with local maxima and right hand side is a location mask image with guidelines

2.4. Individual text line extraction

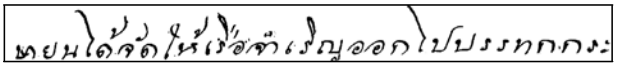
On the location mask text line image, black pixel is found from predetermined line, then contour algorithm is applied to locate the mapping area. Once the specific directions from contour are obtained, other masks in the text line will be ignored so that only the masks of interest are inverted and shown in color on the temporary image. In this process, bit-wise operators are used for mapping the mask text line on the temporary image with text image in order to extract individual text line. Finally, the selected mask text line will be deleted from the source image. The proposed algorithm continually scans through the line for finding the next mask until the end of the line. The algorithm performs the same process for the next line until it finishes with the last line in the mask text line image. Figure 6 shows the process of individual text line extraction.



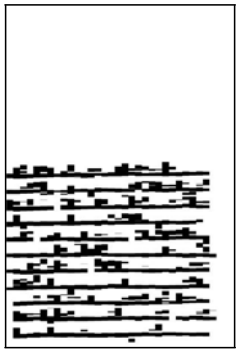
(a)

(b)

(c)



(d)



(e)

Figure 6. Text line extraction step. (a) Mask text line image. (b) Interested mask on temporary image. (c) Text image. (d)Text line image from mapping. (e) Mask text line image with deleted previous line.

3. Experiment

To test the proposed method, 20 pages of document from the King Rama 5 archives from the Nation Archive of Thailand were chosen randomly. These documents were written on blank papers. The documents were scanned at 300dpi and transformed to gray scale images.

The developed method attempts to minimize the step of text line location and extraction. PSL was applied to define the region of text line. Moreover, local maxima projection profile was used to improve the speed for extracting the text lines. To determine the performance of the proposed technique, the numbers of right text line were counted.

Out of the 20 test images, they contain 328 individual text lines in total. After being processed by the proposed approach, there are 17 instances that two lines were combined into one, 3 instances with three lines incorrectly combined, and one case of four lines being incorrectly merged. Hence, the accuracy of the system is therefore considered to be 85.7% based on the following expression.

$$(328 - ((17 \times 2) + (3 \times 3) + (1 \times 4))) / 328 = 85.7\%.$$

In the experiment, the errors of line extraction are due to the connecting of word level between two text lines. Figure 7 shows the error of the inability to separate the two text lines from the extraction process.

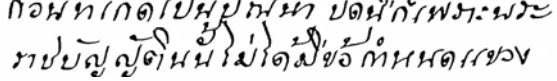


Figure 7. Error due to inability to extract the two text lines

4. Conclusion

In this paper, a new approach for the extraction of text lines from handwritten Thai documents is presented. The method applied the concept of PSL performing as ALCM in order to build mask text lines. Moreover, local maxima profile is utilised for enhancing the speed of extraction. Initial results of this work yielded an accuracy of 85.7% and further experimentation and improvement of the technique will be continued.

References

[1] Pavlidis T. and J. Zhou, "Page segmentation by white streams", Proc. 1st Int. Conf. Document Analysis and Recognition (ICDAR),Int. Assoc. Pattern Recognition, pp. 945–953, 1991.

[2] Ciardiello G., G. Scafuro, M. T. Degrandi, M. R. Spada, and M. P. Roccotelli, "An experimental system for office document handling and text recognition", Proc 9th Int. Conf. on Pattern Recognition, pp. 739–743, 1988.

[3] Nagy S. C. S. G. and S. D. Stoddard, "Document analysis with expert system", Procedings of Pattern Recognition in Practice II, June 1985.

[4] Yanikoglu B. A. and P. A. Sandon, "Segmentation of offline cursive handwriting using linear programming", Pattern Recognition, Vol 31, No. 12, pp.1825–1833, 1998.

[5] Kavallieratou E., N. Dromazou, N. Fakotakis, and G. Kokkinakis, "An integrated system for handwritten document image processing", International Journal of Pattern Recognition and Artificial Intelligence, Vol 17, No. 4, pp. 617–636, 2003.

[6] Senior A. W. and A. J. Robinson, "An off-line cursive handwriting recognition system", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 20, No. 3, pp. 309–321, March 1998.

[7] Manmatha R. and J. L. Rothfeder, "A scale space approach for automatically segmenting words from historical handwritten documents", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 27, No. 8, pp. 1212–1225, August 2005.

[8] Roy K., U, Pal, and B.B. Chaudhuri, "Neural Network based Wordwise Handwritten Script Identification System for Indian Postal Automation", Intelligent Sensing and Information Processing, pp. 240 - 245, 2005.

[9] Shi Z., S. Setlur, and V. Govindaraju, "Text extraction from gray scale historical document images using adaptive local connectivity map", In 8th International Conference on Document Analysis and Recognition, ICDAR, Seoul, Korea, Vol 2, pp. 794–798, August 2005.

[10] Otsu N., "A threshold selection method from gray-level histograms", IEEE Transactions on Systems, Man, and Cybernetics, SMC, Vol 9, No. 1, pp. 62–66, January 1979.

[11] Seksan S., C.C. Fung, "Using content based image retrieval techniques for the indexing and retrieval of Thai handwritten document", The 2nd International Conference on Computer and Automation Engineering, (ICCAE 2010): Singapore. pp. 98 ʊ 101, 26 -28th February 2010.