SEGMENTATION OF TOUCHING LANNA CHARACTERS

Sakkayaphop Pravesjit and Arit Thammano
Computational Intelligence Laboratory, Faculty of Information Technology, King Mongkut's Institute of Technology Ladkrabang, 10520, Bangkok, Thailand

Keywords: Character segmentation, Touching character, Dissection method.

Abstract: Character segmentation is an important preprocessing step for character recognition. Incorrectly segmented characters are not likely to be correctly recognized. Touching characters is one of the most difficult segmentation cases which arise when handwritten characters are being segmented. Therefore, this paper emphasizes the interest to the segmentation of touching and overlapping characters. In the proposed character segmentation process, the bounding box analysis is initially employed to segment the document image into images of isolated characters and images of touching characters. The thinning algorithm is applied to extract the skeleton of the touching characters. Next, the skeleton of the touching characters is separated into several pieces. Finally, the separated pieces of the touching characters are put back to reconstruct two isolated characters. The proposed algorithm achieves an accuracy of 75.3%.

1 INTRODUCTION

Lanna language was used in the 13th to 18th centuries in the Kingdom of Lanna. However, after the kingdom had been annexed by Siam (as Thailand was called until 1939) in 1774, Lanna script became obsolete and was replaced with Thai script. Few people nowadays know how to read or write this language. Lanna manuscripts were typically written about the people's ways of life, believes, laws, folklore, herbal medicine ingredients, history, astrological knowledge, and other general knowledge. Lanna manuscripts were generally inscribed on palm leaves (Figure 1), or on the surface of stones. As time goes by, these ancient documents have been decayed, damaged, destroyed, or lost. Termites, insects, and general decay have left these old manuscripts in poor condition (Figure 2). In order to preserve valuable historical information inscribed on these documents, computerized systems must be put to use in order to translate the inscribed script into the current Thai script.

Touching and overlapping of characters is the first problem encountered when attempting to recognize the handwritten Lanna characters. Touching of characters can emerge when two or more adjacent characters are written too close; therefore, some parts of characters are connected (Saba, Sulong, and Rehman, 2010).

Figure 1: Palm leaf manuscripts.

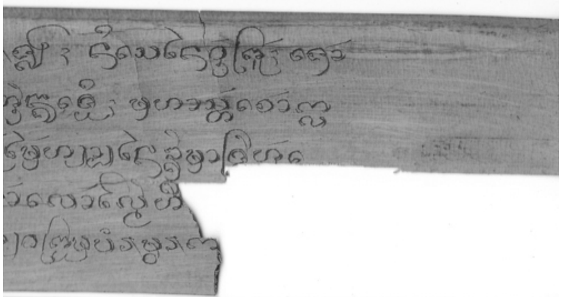
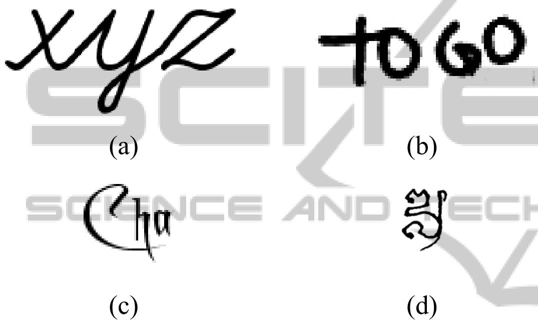


Figure 2: Example of Lanna script.

There are many kinds of touching characters commonly found in the written documents (Figure 3). This is because characters of each language have

Pravesjit S. and Thammano A..
SEGMENTATION OF TOUCHING LANNA CHARACTERS.
DOI: 10.5220/0003511300470051
In *Proceedings of the International Conference on Signal Processing and Multimedia Applications* (SIGMAP-2011), pages 47-51
ISBN: 978-989-8425-72-0

different styles and characteristics. Therefore, the types of touching characters vary from language to language, which in turn require different methods for segmenting the touching characters in each language. For example, the handwritten cursive characters shown in Figure 3(a) are a type of touching characters typically found in English handwritten manuscripts but not in Lanna manuscripts. Only the types shown in Figures 3(b), 3(c), and 3(d) can be found in Lanna manuscripts. The purpose of this research is to separate the touching or overlapping Lanna characters, which have not been effectively solved using any other character segmentation methods.


(a) (b)

(c) (d)

Figure 3: Four types of touching characters.

Character segmentation is a process that seeks to decompose a sequence of characters into individual symbols. There have been substantial researches undertaken to solve character segmentation problem, mostly for numerals, English script, Chinese script, Arabic script, and Bangla script. Segmentation strategies can be divided into three main categories (Casey and Lecolinet, 1996; Marinai, Gori, and Soda, 2005): dissection methods, recognition-based methods, and holistic methods. Dissection methods decompose the image into a sequence of sub-images using general features, e.g., character height and width (Hoang, Tabbone, and Pham, 2009). Recognition-based methods search the image for components that match classes in its alphabet. Holistic methods seek to recognize entire words as a whole, thus avoiding the need to segment the image into characters. Among the methods proposed for character segmentation, Tseng and Chen (1998) proposed a three-stage Chinese character segmentation algorithm. Firstly, a bounding box is created around each stroke of a Chinese character. Secondly, the knowledge-based merging operations are used to merge the stroke bounding boxes together. Finally, a dynamic programming is used to find the optimal segmentation boundaries. The experimental results show that the proposed algorithm is a very effective segmentation algorithm. It works well even with touching and/or overlapping characters. Xiao and Leedham (2000) proposed a novel approach to English cursive script segmentation. In the proposed approach, connected components are split into sub-components based on their face-up or face-down background regions. Then the over-segmented sub-components are merged into characters according to the knowledge of character structures are their joining characteristics. Bhowmik, Roy, and Roy (2005) proposed a segmentation scheme for handwritten Bangla words. The authors use the analysis of directional chaincode and the positional information to extract the features from the image, then employ multilayer perceptron neural network to determine the segmentation points. The authors also point out that their segmentation result can be significantly improved if their proposed technique is combined with the recognition process in a holistic system.

This study focuses mainly on the dissection methods. Projection analysis, connected component processing, and bounding box analysis are three widely used dissection methods (Chen, Wu, and Lee, 1998). While projection analysis is very effective for segmenting good quality machine printed manuscripts (Casey and Lecolinet, 1996), it has limited success when segmenting handwritten manuscripts. Connected component processing and bounding box analysis usually offer an efficient way to segment handwritten manuscripts. However, they might lead to incorrect segmentation when dealing with touching characters as shown in Figure 4.

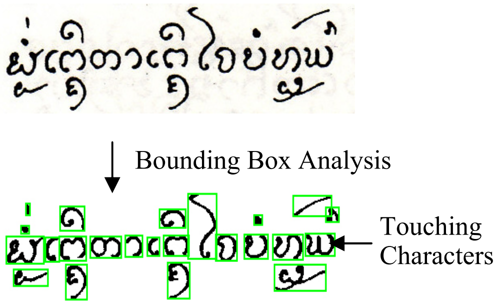


Figure 4: Segmentation results by bounding box analysis.

In this paper, the new dissection algorithm is proposed to segment touching Lanna characters. The performance of the proposed algorithm is measured by the ability of the proposed algorithm to correctly segment 6 different handwritten Lanna manuscripts.

Following this introduction, section 2 briefly describes the general process of the proposed character segmentation process. Section 3 explains the proposed character segmentation algorithm. In section 4, the experimental results are presented and discussed. Finally, section 5 is the conclusion.

2 METHODOLOGY

The process of this Lanna character segmentation has been divided into five steps:

(1) Scan a manuscript and convert it into a binary image.

(2) Use the bounding box analysis to segment the entire manuscript image into subimages. Since the bounding box analysis is only effective in segmenting nontouching characters, the obtained subimages therefore consist of both images of isolated characters and images of touching characters.

(3) Detect touching characters by looking at the aspect ratio (width/height) of each subimage. From the facts that (1) touching characters commonly have an aspect ratio larger than single isolated characters and (2) the aspect ratio of Lanna isolated characters is typically smaller than 3/2, therefore, if the aspect ratio of the subimage is larger than 3/2, it will be identified as the touching characters.

(4) Use the thinning algorithm to reduce the thickness of the character image to its skeleton, which is then sent to the segmentation engine.

(5) Employ the proposed character segmentation algorithm to decompose the touching characters into individual characters.

3 THE PROPOSED CHARACTER SEGMENTATION ALGORITHM

A description of the proposed segmentation algorithm is given below.

A. Search the touching characters image for end points and junction points. Then, adopting the junction points as the partition points, break up the touching characters into several pieces. From each end point, start tracing the contour of touching characters until the nearest junction point is reached. Then, extract such contour from the touching characters. For example, in Figure 5, four end points and two junction points are found. By using the above mentioned technique, the touching characters are broken into 5 contours as shown in Figure 6. Out of the five contours, four contains both a junction point and an end point while one contains two junction points (with no end point on the contour). The contour which contains no end point is a part of the touching characters where two characters touch each other.

B. For each extracted contour, translate the contour so that the junction point coincides with the origin (where x and y axes intersect).

$$T\begin{bmatrix} C_{ix} \\ C_{iy} \end{bmatrix} = \begin{bmatrix} C_{ix} \\ C_{iy} \end{bmatrix} - \begin{bmatrix} J_{ix} \\ J_{iy} \end{bmatrix} \quad (1)$$

where T is the translation operator.

C_{ix} and C_{iy} are x and y coordinate of the i^th contour

J_{ix} and J_{iy} are the x and y coordinate of the junction point of the i^th contour.

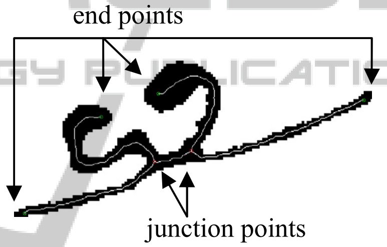


Figure 5: Example of touching characters.

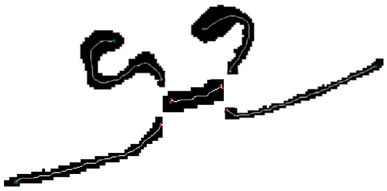


Figure 6: Five contours of the example touching characters.

C. At the origin, create the reference unit vectors along the x axis ((1, 0) and (-1, 0)) and along the y axis ((0, 1) and (0, -1)).

D. For each translated contour, create a vector V_1 that starts at the origin and ends one pixel away from the starting point. Then, determine an angle between the vector V_1 and the x axis in a clockwise direction by using the following equation:

$$\theta_n = \cos^{-1} \frac{U \cdot V_n}{\|U\|\|V_n\|} \quad (2)$$

where$U$ is a unit vector along the x axis. V_n is the nth vector along the translated contour i.

Next, sequentially create the vectors V_2, V_3, …, V_n, …, V_N. However this time instead of starting the vector at the origin, start the vector V_n at the point where the vector V_{n-1} ends. For example, the starting point of the vector V_2 is the ending point of the vector V_1. After each vector V_n is created, determine the clockwise angle of the vector V_n relative to the x axis. If the angle of three consecutive vectors is the same, stop the creation of further vectors.

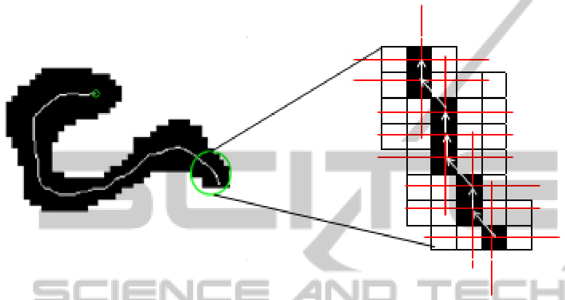


Figure 7: Illustration of step D.

E. Use the linear regression method to calculate the equation of the best-fit line for a series of points, starting at the origin and ending at the ending point of the third consecutive vector. Calculate the angle of the line from the overlapping contour line and assign it to represent the angle of the whole contour.

$$y = a + bx \tag{3}$$

$$b = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2} \tag{4}$$

$$a = y - bx \tag{5}$$

F. After obtaining the angles of all contours, compare the angle of each contour to that of other contours. According to the characteristic of Lanna characters whose trajectories typically do not abruptly change the direction, group the closest match, angle wise, together. Finally, the contours within the same group are combined to form each isolated character.

4 EXPERIMENTAL RESULTS

In this study, the images of Lanna characters used in testing the performance of the proposed algorithm were obtained from 6 manuscripts. Samples of the tested manuscripts are shown in Figure 8. Each manuscript is written by a different handwriting script. Four pages from each manuscript were scanned and processed using steps 1 through 5 outlined in section 2. In the scanned pages, a total of 85 touching characters were found. All of them only consist of two characters. Similar to other languages, touching characters of three or more components are very uncommon in Lanna manuscripts.

Of the 85 touching character images, 75.3% were correctly segmented with the proposed algorithm. Figure 9 shows some of correctly and incorrectly segmented Lanna characters.

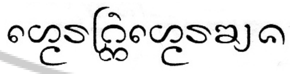
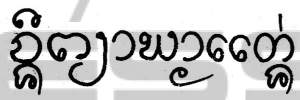
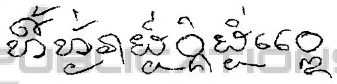
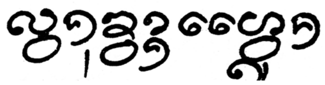
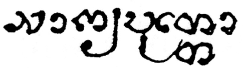
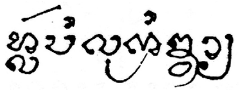
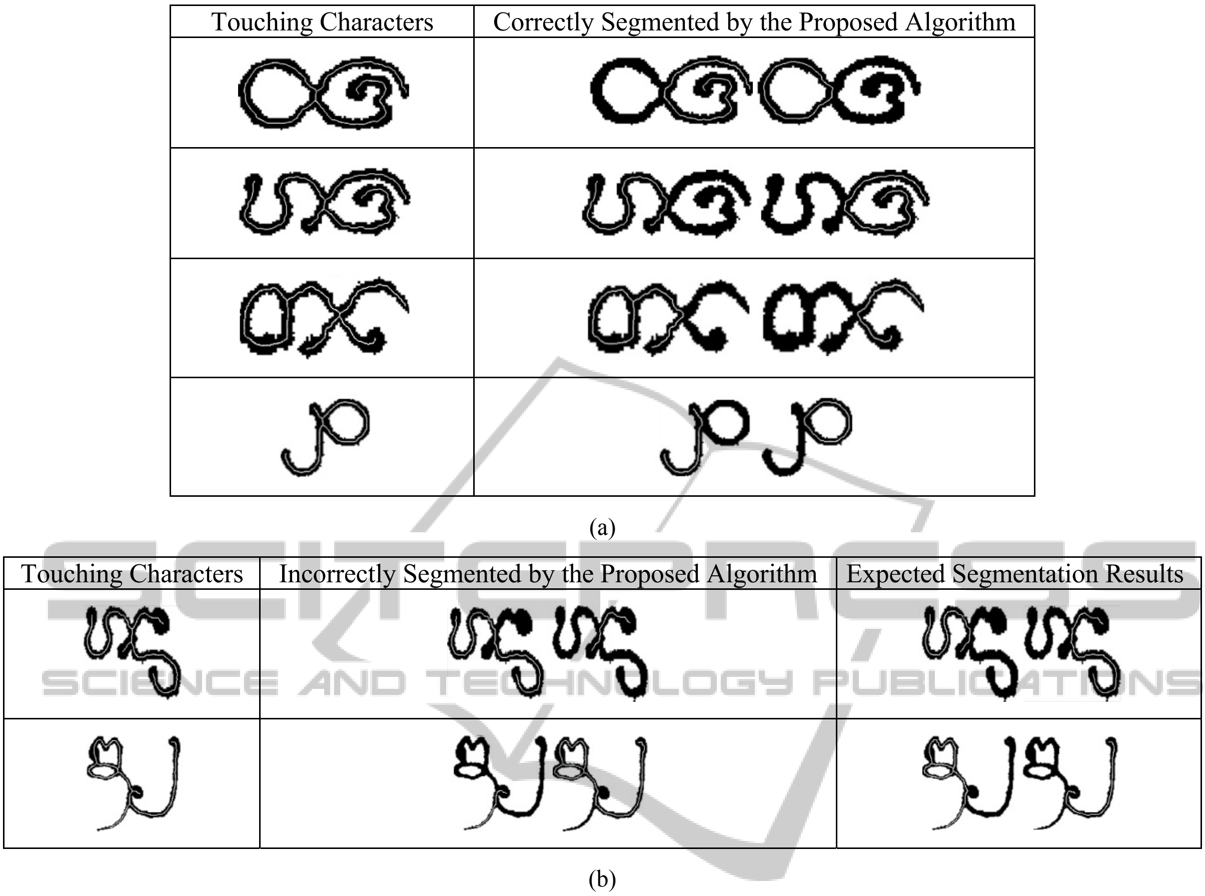
Manuscript 1	
Manuscript 2	
Manuscript 3	
Manuscript 4	
Manuscript 5	
Manuscript 6	

Figure 8: Samples of the tested manuscripts.

5 CONCLUSIONS

A new segmentation algorithm for the off-line handwritten Lanna characters is proposed in this paper. To segment real world documents, segmentation of touching characters is a major problem we have to deal with. With the use of the characteristic of Lanna characters, the touching characters is separated into several pieces, then the separated pieces of the touching characters are put back to reconstruct two isolated characters. From the experimental results, it is clear that the proposed algorithm is quite capable of segmenting touching Lanna characters.


Touching Characters
Correctly Segmented by the Proposed Algorithm
Touching Characters
Incorrectly Segmented by the Proposed Algorithm
Expected Segmentation Results

(a)

(b)

Figure 9: Some segmentation results, where (a) is correctly segmented and (b) is incorrectly segmented.

REFERENCES

Bhowmik, T. K., Roy, A., Roy, U., 2005. Character Segmentation for Handwritten Bangla Words Using Artificial Neural Network. In: *Proceedings of the International Workshop on Neural Networks and Learning in Document Analysis and Recognition*.

Casey, R. G., Lecolinet, E., 1996. A Survey of Methods and Strategies in Character Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 7, pp. 690-706.

Chen, J. L., Wu, C. H., Lee, H. J., 1998. Chinese Handwritten Character Segmentation in Form Documents. *Document Analysis Systems: Theory and Practice*, LNCS 1655, pp. 348-362.

Hoang, T. V., Tabbone, S., Pham, N., 2009. Recognition-based Segmentation of Nom Characters from Body Text Regions of Stele Images Using Area Voronoi Diagram. In: *Proceedings of the 13th International Conference on Computer Analysis of Images and Patterns*.

Marinai, S., Gori, M., Soda, G., 2005. Artificial Neural Networks for Document Analysis and Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 1, pp. 23-35.

Soba, T., Sulong, G., Rehman, A., 2010. A Survey on Methods and Strategies on Touched Characters Segmentation. *International Journal of Research and Reviews in Computer Science*, vol. 1, no. 2, pp. 103-114.

Tseng, L. Y., Chen, R. C., 1998. Segmenting Handwritten Chinese Characters Based on Heuristic Merging of Stroke Bounding Boxes and Dynamic Programming. *Pattern Recognition Letter*, vol. 19, pp. 963-973.

Xiao, X., Leedham, G., 2000. Knowledge-based English Cursive Script Segmentation. *Pattern Recognition Letters*, vol. 21, pp. 945-954.