Segmentation of Thai Handwritten Text for Automatic Document Retrieval

Rapeeporn Chamchong [1], Chun Che Fung [2]

School of Information Technology, Murdoch University
Murdoch, WA 6150 Australia
[1] rapeeporn.c@gmail.com
[2] l.fung@murdoch.edu.au

***Abstract*—There is a huge amount of documents in Thai government organizations. Although automatic document image retrieval systems in English have been proposed and developed, there are no specific system which is capable to retrieve relevant information from documents in Thai language. While matching words or optical character recognition (OCR) can be applied, segmentation of the words and characters is essential to separate them in the first place. There are also both printed and handwritten characters in Thai government documents which pose an additional challenge. While the printed texts can be segmented easily using classical approach, handwritten scripts are hard to separate. The objective of this paper is to present a survey of the existing methods which have been developed recently, and the segmentation techniques of document images in handling Thai printed and handwritten scripts.**

I. Introduction

Currently, computer technology can store and process large amount of image documents in multimedia systems. Therefore, it is theoretically possible to collect and access documents from computer systems within government departments. In the case of Thailand, there are many types of official documents, such as memorandums, announcements, letters, government gazette, circulars in various forms and records. However, Thai government documents include not only printed texts, but also handwritten text, signatures, symbols, drawings, paintings and other relevant information. It is therefore necessary to store Thai government documents as images to preserve their integrity. Although the current storage systems are capable to store document images, there is no specific system which is able to retrieve relevant information and knowledge from these documents. Accordingly, a system which can efficiently retrieve data from document images based on image processing technique with an ability to retrieve relevant knowledge from the documents will be highly desirable. The motivation of this study is to conduct a survey on the state-of-the-art techniques for the problem of processing and retrieval of Thai handwritten text.

One obvious solution for document image retrieval is by matching directly the image data using images of the words as queries or optical character recognition (OCR). Word and character segmentation is generally seen as a preprocessing step for tasks such as document structure extraction, printed character or handwriting recognition.

As mentioned earlier, there exist both printed and handwritten characters in Thai government documents. Comparatively speaking, the printed scripts can be separated easily using mechanism such as OCR. Despite much research, handwritten scripts still pose as an academic challenge as they are difficult to segment. Handwritten pages may include narrow spaced lines with overlapping and touching components, incomplete writings and blurred images are just a few of the issues involved. In addition, characters and words have unusual and varying shapes, and they are writer-dependent. Other issues are the periods and their places. The Thai vocabulary is also very large and may include proper and unusual names and words. Full text recognition is therefore in most cases not yet available, except for printed documents for which dedicated OCR have been developed.

In terms of document analysis, it is desirable to have embedded tools for searching of blocks, lines and words, and the inclusion of a dedicated handwriting recognition system. Interactive tools are generally offered for segmentation and recognition correction purposes. Several projects in the past are concerned with printed materials. However, solutions to tackle handwritten text perfectly are yet to be developed.

The purpose of this article is to survey the research efforts made for Thai government documents on the text line, word and character segmentation task. Section 2 describes the properties of Thai language and section 3 explains about text line, word and character segmentation. Section 4 is a description of research in handling Thai printed and handwritten text segmentation. Finally, discuss and suggestions for future research are given in the last section.

II. Properties of Thai Language

A. Thai Alphabet Set

TABLE I
THAI ALPHABET SET

Type	Elements
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 ล ว ศ ษ ส ห ฬ อ ฮ
Vowels	ะ า ิ ี ึ ื ุ ู ั ำ เ แ โ ใ ไ ฤ ฦ
Tones	่ ้ ๊ ๋
Symbols	็ ์ ๆ ฯ

The Thai character set [1] consists of 44 consonants, 17 vowels, 4 voice tones, and 4 special symbols, as shown in

Table I. Chatchinarat's research [2] divides consonants into two groups in order to perform connected identification.

TABLE II
THAI CONSONANT GROUPS

Consonants	Elements
Group 1	ก ข ข ค ต ฆ ง จ ฉ ช ซ ฌ ญ ฏ ฏ ฐ ฑ ฒ ณ ด ต ถ ท ธ น บ ป ผ ฝ พ ฟ ภ ม ย ร ล ว ศ ษ ส ห ฬ อ ฮ
Group 2	ฌ ญ ฎ ฏ ฒ ณ ฬ

The character in group 2 can be considered as one character or two linkage characters as show in Fig. 1

Fig. 1. Consonant Characters of Thai writing

B. Thai Writing System

Thai writing starts from left to right and from top to bottom. Thai writing does not require spaces between words and sentences might be separated by spaces or might be not. There are four levels of writing [3], which are the tone, upper vowel, body, and lower vowel levels, as shown in Fig. 2.

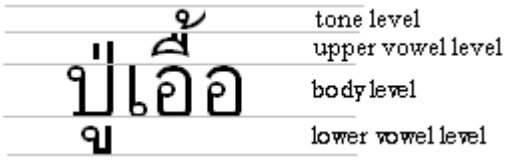

Fig. 2. Four levels of Thai writing

The tone level is the top position in a line followed by the upper vowel, the body, and the lower vowel level. The consonants are located at the body level and the tone marks are always located at the tone level. The other characters have their specific positions. The position level can be used to help to identify the character.

III. TEXT LINE, WORD AND CHARACTER SEGMENTATION

Before the characters are segmented, text lines should be segmented. This can be processed by histogram projection in horizontal. However, some problems may occur in horizontal projection processing. For instance, text lines may be slanted or some characters may be overlapped. This problem can be solved by rotating the lines or the document. In segmentation process, printed scripts and handwritten scripts have different styles. It is found that printed scripts have equal gap between every characters. Consequently, printed texts can be separated easily. In addition, it is found that English language differs from Thai language. Word of English can be separated by space and stop sentence by full stop. Thus, segmentation of English scripts for document retrieval is easier than Thai scripts. Words in English texts can be segmented by recognizing the spaces and the sentences can be segmented by identifying the full stops. Then, the words and sentences can be used to generate the indexes for matching the words and sentences in the document. There are several techniques [4-8] which have been developed to separate words and sentences in English.

The process of character segmentation consists of three steps. Text line segmentation is the first step. This is followed by word segmentation and then character segmentation.

A. Text Line Segmentation

The first step of character segmentation is text line extraction. This step has been surveyed comprehensively by Laurence and et al. in [9]. Several techniques in line segmentations are:

- Projection-based methods
- Smearing methods
- Grouping methods
- Methods based on the Hough transform
- Repulsive-Attractive network method
- Stochastic methods

Most of these methods are capable to extract lines. The projection, smearing and Hough-based method are classical approaches applied to straight lines and they can be enriched by local considerations. The stochastic methods can avoid overlapping components. If text line fluctuations, smearing and grouping methods are convenient. The recurrent nature of the repulsive-attractive method suits for similar lengths of lines.

B. Word Segmentation

After the lines are extracted, English, Latin, Indian, Arabic scripts can be separated. However, other scripts such as Thai cannot separate the words easily as there are no stop words like English language. Hence, word segmentation techniques [4-6, 8, 10] have been proposed by a variety of researchers. For example, vertical profile projection, scale space technique [11], word spotting[8], and word-wise technique[10] are commonly used. It is found that word segmentation techniques avoid the problem of linkage's characters or character segmentation. These techniques, which are known as "holistic methods" [12], suit for document retrieval. Unfortunately, there are problems in Thai scripts thereby making these methods unsuitable.

C. Character Segmentation

Character segmentation has been proposed many years ago. R. G. Casey and E. Leolinet [12] have reported a survey of the methods and strategies in character segmentation. The four strategies for segmentation are listed below.

1) The classical approach: in which segments are identified based on "character-like" properties such as height, width, separation from neighbouring components, disposition along a baseline, etc. This technique cuts image into a sequence of sub-images, which are called "dissections". The popular methods of these are "white space and pitch", "projection analysis", "connect component processing" and "dissection with contextual post-processing graphemes"

2) Recognition-based segmentation: in which the system seeks the image for components that match classes in the alphabet. This technique can be divided into two methods,

which are “windowing process” and “feature-based”. The windowing process can operate directly on the image pixels or applying in the form of weightings or grouping of positional feature measurements made on the images. The feature-based can be performed by Hidden Markov Model (HMM) and Non-Markov approaches (such as N-gram technique).

3) Holistic methods: in which the system searches to recognize whole words, hence, avoiding separation of character. This method performs feature extraction in the first step, then global recognition by comparing the representation of the unknown word with those of the references stored in the lexicon. Consequently, this method uses the “classical approach”, with complete words as the symbols to be recognized. For instance, scale space technique[5] and holistic word recognition [13] are normally used.

4) Hybrid approaches: in which the combination of first three strategies by weighting. This effectively is an integration of the previous techniques and the weights can be decided by heuristics or subjectively.

In the course of this study, it was further found that the various researches of Thai printed character segmentation always use dissection techniques [14] due to its robustness and appropriateness.

IV. THAI HANDWRITTEN SEGMENTATION

There are several problems of Thai handwritten segmentation. The key ones being the different individual styles and there is not stop word like the English language. This is a main problem of Thai writing system; therefore it is difficult to separate the words or sentences. In addition, there are four levels of Thai writing system so it is essential to segment the line levels to identify the type of character. Moreover, Thai words are written as non-cursive scripts, however, overlapping of several parts of the characters can always be found. However, there are only few researches in Thai handwritten segmentation.

Among the papers concerning Thai handwritten segmentation, they always use the classical approaches to segment character [1, 3, 14-16] and a lot demonstrations are based on controlling the writing styles of the writers.

In M. Lohakan and et al. [17], they proposed the single character segmentation of handwriting approach by using patterns of character levels to identify different type of character. In the experimental results, 100 examples of input written by 10 subjects were implemented. The segmentation accuracy achieved was 98.0%.

The adaptive block approach to segment character was proposed by S. Arunrungrusmi and K. Chamnongthai [15] . The blocks can be formed by utilizing the distinctive features of Thai characters such as cavity, local minimum and maximum points. The experiments are based on a collection of data sets of Thai handwriting from ten Thais. The results of the recognized character can then be applied as a part of the reorganization system. However, some errors were found to remain due to scrawling characters and losing of the distinctive features.

In addition, the water reservoir technique, proposed by U. Pal [11] for Indian scripts is another premising approach. This technique is applied by A. Chatchinarat and B. Kruatachue [2] for online Thai handwritten segmentation. However, this research was performed on only two or three characters. In the experiment, a set of data 200 isolated and 200 connected components were collected from Thai handwritten text. Isolated detection has accuracy 98%. Connect detection has achieved an accuracy of 95%. Overall detection has an accuracy of 96.5%. Most of the errors come from connected characters which are identified as isolated characters.

In O. Surinta and R. Chamchong’s work [18], they applied the recursive strip for line segmentation and the classical approach for character segmentation by using connect component from Thai palm leaves. In the experiment, the number of data set of palm leaves was 5 binds, which composed of 227 images. The completed images after background elimination are 138 images (60.9%) following text lines extraction. The complete line images are 131 lines (78.5%). Finally, the ancient Thai characters segmented correctly were 71.81%. This experiment found several ancient Thai characters cannot be separated due to overlapped components and connected characters. While it was found that these techniques could be applicable, they however are not useful for real documents.

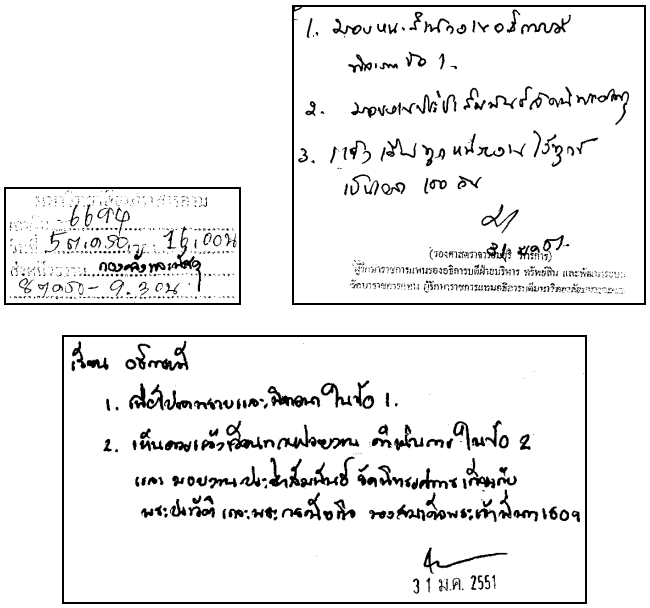

Fig. 3. Examples of Thai writing styles in the government documents

V. CONCLUSION AND DISCUSSION REMARKS

This paper is the initial work of a research project aiming to provide a better solution for document retrieval from Thai government document. In the domain of Thai document image retrieval, character segmentation methods have been developed in several projects which perform transcript mapping authentication, word mapping or word recognition. As the need for recognition and mapping of handwritten material increases, application of character segmentation will be expected to increase rapidly. Contrary to printed text,

handwritten documents have unique characteristics due to the individual writing styles of the writers. There is no specific segmentation method which is suitable for Thai government documents. While the printed text can be easily separated by classical approaches, there are other issues with handwritten scripts in Thai. The problems are due to character size, stroke width, average spacing and connected characters. The organization of segmentation methods has been summarized in this paper. There are three main steps, which are text line, word and character segmentation. The techniques surveyed here have been proposed to separate particular sets of documents. They can be also used for other documents with similar characteristics.

VI. REFERENCES

[1] C. Yingsaeree and A. Kawtrakul, "Rule-based Middle-level Character Detection for Simplifying Thai Document Layout Analysis," in Proceedings of the 2005 Eight International Conference on Document Analysis and Recognition (ICDAR'05), vol. 2, 2005, pp. 888- 892.

[2] A. Chatchinarat and B. Kruatrachue, "Online Thai Handwritten Segmentation Using Water Reservoir Technique," presented at The International MultiConference of Engineers and Computer Scientists 2007 Hong Kong, 2007.

[3] W. Chatwiriya, "Off-line Thai Handwriting Recognition In Legal Amount," in Computer Engineering, vol. Doctor of Philosophy: West Virginia University, 2002, pp. 190.

[4] R. Manmatha, C. Han, and E. M. Riseman, "Word Spotting: A New Approach to Indexing Handwriting," in 1996 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'96). San Francisco, CA, USA, 1996, pp. 631-637.

[5] R. Manmatha and J. L. Rothfeder, "A Scale Space Approach for Automatically Segmenting Word from Historical Handwritten Documents," IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 27, pp. 1212-1225, 2005.

[6] S. Marinai, E. Marino, and G. Soda, "Adaptive Word Indexing of Modern Printed Documents," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, pp. 1187-1199, 2006.

[7] J. Sas and U. Markowska-Kaczmar, "Semi-Supervised Handwritten Word Segmentation Using Character Samples Similarity Maximization and Evolutionary Algorithm " in 2007 6th International Conference on Computer Information Systems and Industrial Management Applications, 2007, pp. 316-321.

[8] K. Zagoris, N. Papamarkos, and C. Chamzas, "Web Document Image Retrieval System Based on Word Spotting," in 2006 IEEE International Conference on Image Processing. Atlanta, GA, 2006.

[9] L. Likforman-Sulem, A. Zahour, and B. Taconet, "Text line segmentation of historical documents: a survey," International Journal on Document Analysis and Recognition vol. 9, pp. 123 - 138, 2007.

[10] K. Roy and U. Pal, "Word-wise Hand-written Script Separation for Indian Postal Automation," in Tenth International Workshop on Frontiers in Handwriting Recognition (2006), 2006.

[11] N. Tripathy and U. Pal, "Handwriting Segmentation of Unconstrained Oriya Text," in Ninth International Workshop on Frontiers in Handwriting Recognition (IWFHR'04), 2005, pp. 306-311.

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

[13] V. Lavrenko, T. M. Rath, and R. Manmatha, "Holistic Word Recognition for Handwritten Historical Documents," in The Proceedings of Document Image Analysis for Libraries (DIAL) 2004, 2004.

[14] N. Premchaiswadi, W. Premchaiwadi, S. Duangphasuk, and S. Narita, "SMILE: Printed Thai Character Recognition System," WSEAS Transactions on Computers, vol. 2, pp. 430-434, 2003.

[15] S. Arunrungrusmi and K. Chamnongthai, "Adaptive Blocks for Skeleton Segmentation in Handwritten Thai Character," in The 2000 IEEE Asia-Pacific Conference on Circuits and Systems, 2000. (IEEE APCCAS 2000) Tianjin, China, 2000, pp. 767-770.

[16] S. Watcharabutsarakham, "Using Projection and Loop for Segmentation of Touching Thai Typewritten," in IEEE International Symposium on Communications and Information Technology, 2004 (ISCIT 2004) vol. 1, 2004, pp. 504- 508

[17] M. Lohakan, S. Airphaiboon, and M. Sangworasil, "Single-character Segmentation for Handprinted Thai Word," presented at Fifth International Conference on Document Analysis and Recognition (ICDAR'99), 1999.

[18] O. Surinta and R. Chamchong, "Image Segmentation of Historical Handwriting from Palm Leaf Manuscripts," presented at IFIP International Federation for Information Processing, China, 2008.