Using Content Based Image Retrieval Techniques for the Indexing and Retrieval of Thai Handwritten Documents

Seksan Sangsawad
School of Information Technology
Murdoch University
Perth, Australia
seksan.s@gmail.com

Chun Che Fung
School of Information Technology
Murdoch University
Perth, Australia
l.fung@murdoch.edu.au

Abstract—**This paper proposes the use of content base image retrieval (CBIR) techniques for indexing and retrieval of handwritten documents in Thai language. Issues associated with Thai handwritten documents are the lack of spacing between words, multi-level alphabets and different writing styles. This causes low recognition rate based on automated techniques such as Optical Character Recognition (OCR). This paper also examined off-line signature recognition techniques in order to adapt to Thai handwriting system for matching data. The objective of the proposal is to develop a semi-automated method to index and retrieve Thai handwritten documents based on sampled keywords by combining CBIR and signature recognition techniques.**

Keywords-Thai handwritten document; CBIR; sampled keywords; signture recognition

I. INTRODUCTION

Nowadays, a huge number of handwritten documents are still stored and processed in their original forms. The collection and access of these documents are part of the daily operations in business organizations, health services, university faculties, finance, and government departments. Retrieval of the document must be done manually and this requires time and efforts, and this also subjects to issues of mishandling and security. In the worst case, the document could be damaged or even lost during the process. While it is possible to digitize such documents by computer and multimedia technologies, it requires time and resources for data entry.

In the case of Thailand, there are a lot of official document types, such as memorandums, announcements, letters, answer sheets, cheques, circulars in various forms and records. However, Thai documents include not only printed texts, but also handwritten text, signatures, symbols, drawings, paintings and other relevant information. It is necessary to store such documents as images to preserve their integrity. While there are multi-media database systems which are capable to store and retrieve the documents, such information have to be input manually and it requires substantial amount of effort. A system that can efficiently retrieve data from document images based on the context and image processing technique with an ability to retrieve relevant knowledge from the documents will be highly desirable.

One of the solutions to address the issue of document retrieval is to use optical character recognition (OCR) to transform the image into text, and then index the document based on its content. However, there are limitations on the accuracy of handwriting recognition applications. For instant, such system can only recognize numeral, or regions of interests in the forms or documents [1, 2]. In the case of unstructured documents, the accuracy rate is not sufficiently high.

In such applications, word and character segmentation are generally used as a preprocessing step for tasks such as document structure extraction, printed character or handwriting recognition. Another solution is matching directly the image data using images as the keywords to query. This is the approach being proposed in this paper.

In addressing the issues related to both printed and handwritten characters in Thai documents, 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 issue is the nature of Thai writing system, each word in a sentence is connected. It is therefore difficult to do segmentation and extraction. 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.

There are many successful OCR products which can do segmentation and extraction almost completely for English printed scripts or other languages that have space to separate each word in sentence. Apart from the fact that Thai words in a sentence is connected, there is the issue of multiple word levels for Thai writings. This is illustrated in Figure 1 and this causes difficulties for OCR to perform segmentation and extraction for Thai languages. A comparison of two

978-1-4244-5586-7/10/$26.00 ©2010 IEEE

handwritten documents in Thai and English is shown in Figure 2. Nevertheless, applications of offline handwriting recognition have been used successfully for the classification of zip code on mails and verifying signatures on cheques [3].

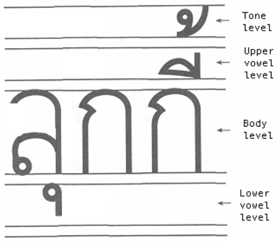
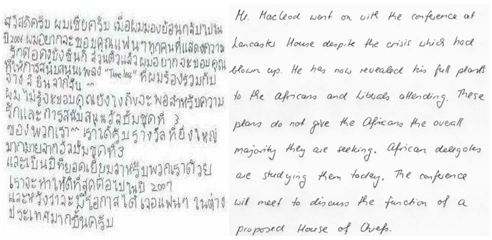


Figure 1. Multiple levels in Thai words



สวัสดีครับ ผมเชียครับ เมื่อผมมองย้อนกลับไปใน
ปี 2006 ผมอยากจะขอบคุณแฟนๆทุกคนที่แสดงความ
รักต่อดงบังชินกิ ส่วนตัวแล้ว ผมอยากจะขอบคุณ
ที่ให้การสนับสนุนเพลง "Time less" ที่ผมร้องร่วมกับ
จาง ลี ยิน จากจีน ^^
ผมไม่รู้จะขอบคุณยังไงถึงจะพอสำหรับความ
รักและการสนับสนุนอัลบั้มชุดที่ 3
ของพวกเรา^^ เราได้รับรางวัลที่ยิ่งใหญ่
มากมายจากอัลบั้มชุดที่3
และเป็นปีที่ยอดเยี่ยมสำหรับพวกเราด้วย
เราจะทำให้ดีที่สุดต่อไปในปี 2007
และหวังว่าจะมีโอกาสได้เจอแฟนๆ ในต่าง
ประเทศมากขึ้นครับ

Mr. Macleod went on with the conference at
Lancaster House despite the crisis which had
blown up. He has now revealed his full plans
to the Africans and Liberals attending. These
plans do not give the Africans the overall
majority they are seeking. African delegates
are studying them today. The conference
will meet to discuss the function of a
proposed House of Chiefs.

Figure 2. Comparing a Thai and an English handwritten document

In terms of document management, the key of information retrieval is the ability to extract and match some desired data. Several tools have emphasized on line and word segmentation. Dictionary techniques are also used for recognition and correction purposes. In the past, several projects have concerned with printed text documents. However, high performance solutions to deal with handwritten document perfectly are not yet to be developed.

This paper describes a proposed method for indexing and retrieving Thai handwritten document images from repository document image database. The method is based on Content Based Image Retrieval (CBIR) and offline signature recognition techniques. In order to retrieve relevant documents, sampled text image or text word are first identified and stored. They are then used as the keywords for subsequent queries. The next section describes some of the current work and it is followed by a description of the proposal. A conclusion is the included in the final section.

II. CURRENT WORKS

Most document retrieval systems have used OCR process for transcribing printed text document images to readable text for subsequent indexing and retrieval. However, there are other techniques that do not need to transcribe the whole document by applying techniques in CBIR systems. For example, QBIC (Query By Image Content) was the first commercial CBIR system for query and retrieval of images from a database. The Virage Search Engine allows querying of still images and video streams, using visual queries based on color, composition (color layout), texture and structure (object boundary information) [4,5]. Another similar system is the Picasso system [6, 7]. This system allows query by global color similarity, color region similarity, color semantics, texture similarity and shape similarity.

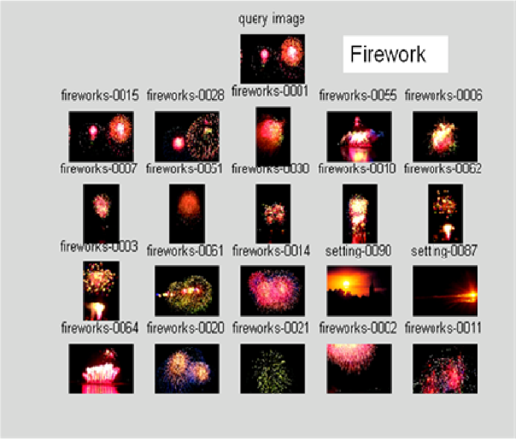


Figure 3. Searching by sample image

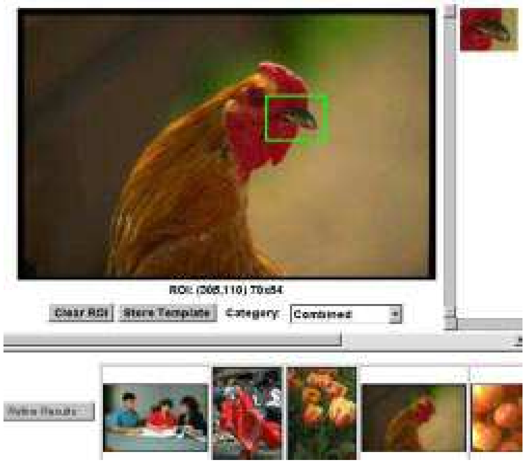


Figure 4. Searching by selected area

Another method for recognition and retrieval of unreadable information is signature recognition techniques. These techniques have included hidden Markov models (HMM) [8], structural techniques, template matching, and feature-based techniques [9]. Feature-based techniques include both local features and global features, or combinations of both. Although local features give a good localized characterization, it is difficult to compute them dependably. Computing global features in sub regions of the signature image can be used to solve this problem. Methods which are based on the characterization of a small set of sub regions of the signature image have been reported in [10]–[13]. Sabourin et al. in [14] described the use of global features that are based on shape matrices for offline signature recognition. In addition, Fang et al. have proposed a method for increasing the amount of training data by generating additional samples from existing ones as described in [15]. There are many other methods that have been developed for signature recognition. In the 1970s, Multiple Neural Network Classification Structures were proposed by Papunzarkos et al. [16]. They made use of global features, grid information features and texture features of the signatures. In the 1980s, Ammar [17] used the statistics of high grey-level pixels to identify pseudo-dynamical characteristics of signatures. The average error rates of this method were 22.8. Stroke positions and positional variations for signature verification were used by Fang et al. [18, 19] with an average error rate of 18.1%. Lizarraga et al. [20] presented personal authentication method using stroke & contour slopes and achieved an error rate of 0.7% and 93.7% correct classification rate.

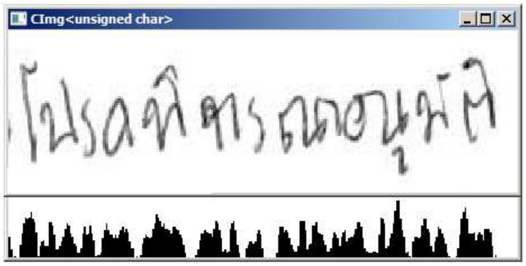
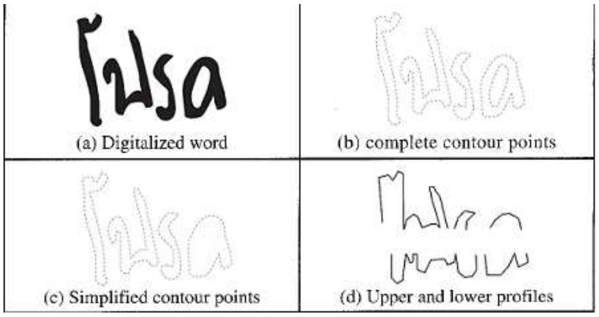


Figure 5. Extraction by histogram


(a) Digitalized word

(b) complete contour points

(c) Simplified contour points

(d) Upper and lower profiles

Figure 6. Extraction by word shape

III. Proposed CBIR Indexing and Retrieval Technique

This paper only focuses on the tasks of retrieval and indexing. Transcription of the whole document is not considered. This paper proposes a method to deal with Thai handwriting documents in form of offline signature recognition. The process starts by scanning the document and then presented to the user. In this step, line segmentation is used, and each line of handwritten is seen as a signature. Feature-based teachnique such as histogram and contour slopes are used. The user is then invited to select portion(s) of the text image in the document and use them as the sampled keyword for subsequent query. Sub region of the image, which is chosen from document, will also be applied with the same feature-based teachnique. They are then used to locate and match similar word(s) in the document. Liner matching can be used to find the positions of relevant word in the document. The system will also automatically scan the relevant documents from the document image database. For the returned documents, user can then check for any correction if necessary. In this step, partial word segmentation are processed. The speed for matching will be increase for subsequent search as the remain handwritten image are organised in smaller sizes. Moreover, during the selection of image keywords and checking for correction, the user can assign tags for indexing and query by text word. Therefore, the system will learn, store the feature vector and rank documents.

As a means of measurement, the frequency counts and computed value of Term-Frequency-Inverse Document Frequency (TF-IDF) weights can be used for counting the number of times a word occurred in a document, and the number of documents in the collection that contains the word. They can also be used to rank the results from the documents or pages, using content-based ranking. This will help to increase speed of retrieval. Finally, precision and recall can be used to evaluate the performance of the system by measuring the relevancy of the retrieved documents and whether all the relevant documents have been retrieved.

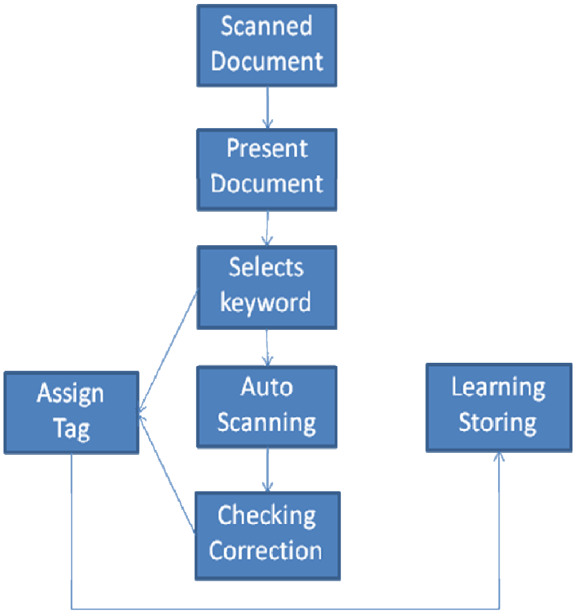


Figure 7. Framework of Proposed CBIR approach for managing hand-written documents

IV. CONCLUSION AND DISCUSSION

Identification and matching of English handwritten documents using micro-feature and macro-feature have already reached an identification rate of 97.94% [21, 22]. One of the key characteristics of English language is the use of space to separate the words in a document. However, Thai language is more complicated and it is difficult to perform world segmentation. This paper has described methods to recognize Thai handwriting in the form of off-line signature recognition that does not need the meaning of the words. Region of interest from CBIR technique can be applied for retrieval in the form of sampled keywords. This proposal is based on a part of the handwritten document and to associate it with some keywords (or tags). This portion of the image (or handwritten text) is then used to search through the other parts of the image and other image files looking for the same sample. Once found, the document will then be tagged or associated with the sampled keywords. Subsequently, the documents could be retrieved or indexed with this sampled keyword. This will provide a more efficient means to access the documents in the future.

Finally, the issues of word segmentation in Thai language are mainly due to the connected words and multiple word levels. Using a sub-region of the image as a query can solve the problem. Partial word segmentation can then be automatically processed, and consequently, matching speed will be improved as the remaining words in the sentences will be cut down to the small size. A design and implementation of the proposal is currently being developed. It is expected that further results will be reported subsequently.

REFERENCES

[1] U. Bohnacker, J. Schacht, T. Yucel, “Matching form lines based on a heuristic search,” Document Analysis and Recognition, vol. 1, 1997, pp. 86 – 90.

[2] W. N. Lin, K. S. Yap, M. Khalid, M. Yusof, “Design of an automated data entry system for hand-filled forms,” TENCON 2000. Proceedings, vol. 1, 2000, pp. 162 – 166.

[3] A. Kornai, K. M. Mohiuddin and S. D. Connell: Recognition of Cursive Writing on Personal Checks. In: Proc. of the 5th Int’l Workshop on Frontiers in Handwriting Recognition. Colchester, UK, September 2-5, 1996.

[4] M. Flicker, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele and P. Yanker, “Query by image and video content: The QBIC system,” IEEE Computer magazine, vol. 28, 1995, pp. 23-32.

[5] J. R. Bach, C. Fuller, A. Gupta, A. Hampapur, B. Horowitz, R. Humphrey and R. Jain, “The Virage image search engine: An open framework for image management,” SPIE storage and Retrieval for Still Image and Video Databases, 1977, pp. 76-87.

[6] A. Del Bimbo, P. Pala, “Retrieval by elastic matching of user sketches,” IEEE Trans. On Pattern Analysis and Machine Intelligence, vol. 19, 1997, pp. 121-132.

[7] A. Del Bimbo, M. Mugnaini, P. Pala, F. Turco, “Visual querying by colour perceptive regions,” Pattern Recognition, vol. 31, 1998, 1241-1253.

[8] J. Coetzer, B. M. Herbst, and J. A. du Preez, “Offline signature verification using the discrete radon transform and a hidden markov model,” EURASIP J. Appl. Signal Process., no. 4, pp. 559–571, 2004.

[9] W. Hou, X. Ye, and K. Wang, “A survey of off-line signature verification,” in Proc. Int. Conf. Intelligent Mechatronics and Automation, 2004, pp. 536–541.

[10] R. Sabourin, G. Genest, and F. J. Preteux, “Off-line signature verification by local granulometric size distributions,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 19, no. 9, pp. 976–988, Sep. 1997.

[11] M. K. Kalera, S. N. Srihari, and A. Xu, “Offline signature verification and identification using distance statistics,” Int. J. Pattern Recognit. Artif. Intell., vol. 18, no. 7, pp. 1339–1360, 2004.

[12] S. N. Srihari, A. Xu, and M. K. Kalera, “Learning strategies and classification methods for off-line signature verification,” in Proc. Int. Workshop Frontiers Handwriting Recognition, 2004, pp. 161–166.

[13] H. Srinivasan, S. N. Srihari, and M. Beal, “Signature verification using kolmogorov-smirnov statistic,” in Proc. Int. Graphonomics Soc. Conf., Salerno, Italy, 2005, pp. 152–156.

[14] R. Sabourin, J.-P. Drouhard, and E. Wah, “Shape matrices as a mixed shape factor for off-line signature verification,” in Proc. ICDAR, 1997, vol. 2, pp. 661–665.

[15] B. Fang, C. Leung, Y. Tang, P. Kwok, K. Tse, and Y. Wong, “Offline signature verification with generated training samples,” in Proc. Inst. Elect. Eng., Vis. Image Signal Process., 2002, vol. 149, pp. 85–90.

[16] N. Papunzarkos, and H. Baltzakis, “Off-Line Signature Verification Using Multiple Neural Network Classification Structures”, IEEE 1977.

[17] Maan Ammar, Yuuji Yoshida, Teruo Fukumura, “Description of Signature Images and its applications to their classification”, IEEE 1988.

[18] B. Fang, C.H. Leung, Y.Y. Tang, K.W. Tse, P.C.K. Kwok, Y.K. Wong, “Offline signature verification by the tracking of feature and stroke positions”, Pattern Recognition Society (2002).

[19] B. Fang, “Tracking of feature and stroke positions for off-line signature verification”, IEEE ICIP 2002.

[20] Miguel G. Lizárraga and Lee L. Ling, “Biometric Personal Authentication Based on Handwritten Signals”, ICBA 2004, LNCS 3072, pp. 533-539

[21] S. N. Srihari, S. Lee, ”Automatic handwriting recognition and writer matching on anthrax-related handwritten mail,” Frontiers in Handwriting Recognition, 2002, pp. 280-284

[22] S. N. Srihari, Z. Shi, “Forensic handwritten document retrieval system,” Document Image Analysis for Libraries, 2004, pp. 188-194.