

WORD RETRIEVAL FROM OLD LANNA MACHINE-PRINTED DOCUMENT BY MATCHING KEYWORD IMAGE

Wilawan Yathongkhum^{1*}, Natsima Suradet², and Jeerayut Chaijaruwanich³

Received: January 16, 2017; Revised: June 23, 2017; Accepted: June 23, 2017

Abstract

In this paper, we propose a method for word retrieval from old Lanna machine-printed document images based on keyword image matching. The research aims at detecting given keywords given by the user and locates it them on document images. The first step successfully creates a synthetic keyword image. Then, we extract a list of candidate words from the document image by using three 3 extraction methodologies: pixel-based sliding window template matching, block-based sliding window template matching, and word feature matching. Next, a feature vector of each candidate word image is extracted using a window-based feature extraction. Finally, all relevant words are retrieved by comparing the similarity between the keyword and the word image feature vector. The similarity between two 2 feature vectors is evaluated using two 2 methods: sub-window similarity and Euclidean distance. The experimental results show that the method using a combination of word feature matching and Euclidean distance provides the best performance. It is shown that the proposed method is feasible, valid, and effective for Lanna word image searching.

Keywords: Lanna document images, word retrieval, keyword image matching, synthetic Lanna word image, Lanna word searching

Introduction

The Lanna language, commonly referred to as Kam Muang is still spoken by local people in Northern Thailand. The Lanna script or Tua Muang is a descendant of the Old Mon scripts including the Lao religious scripts and the Burmese script (Prongthura, 1982). Nowadays, the Lanna script has become obsolete and has been replaced with the Thai script for everyday use of both Thai and Lanna texts. The traditional

alphabet is now mainly limited to texts in Buddhist temples, where many old sermon manuscripts are still in active use. Nonetheless, the Lanna script can still be found in old religious manuscripts, texts, and temple scriptures (Maneechedta, 1994; Pansook, 2008). The Lanna historical documents contain a vast amount of valuable information. Unfortunately, these have degraded from wear and tear, dirt,

¹ Ban Mae Thoei School, Lamphun, 51110, Thailand. E-mail: y.wilawan@gmail.com

² Department of Computer Engineering, Faculty of Engineering, Rajamangala University of Technology Lanna Tak, Tak, 63000, Thailand.

³ Department of Computer Science, Faculty of Science, Chiang Mai University, Chiang Mai, 50202, Thailand. E-mail: y.wilawan@gmail.com

* Corresponding author

artifacts, and misuse, and some of these them have been misplaced and are difficult to locate. Now there are organizations that attempt to preserve the original documents and also keep track of them. To avoid further degradation, paper copies and scanned images have been taken to preserve these resources from further loss and degradation. However, searching through this growing resource of historical information for texts of interest requires scholars who are literate in the Lanna script. This manual translation process has its shortcomings, including the need for trained scholars and that the fact that manual translation can take a lot of time.

Document images are electronic text-based images stored in the form of pixels. The aim of document image retrieval is to enable users to accurately and easily browse or find the relevant information in a massive document image database. In the area of document image retrieval, different methods for searching document images with a keyword have been developed. These methods can be classified into Optical Character Recognition (OCR)-based or image-based (keyword spotting or keyword detection). An OCR-based method uses a text-to-text matching approach. It transforms the document image into a machine-readable form by applying an appropriate recognition process and then retrieval by examining the document with a keyword (Marukawa *et al.*, 1997; Senda *et al.*, 1993). However, this limits the accuracy of the text search and document retrieval tasks due to OCR errors. These OCR errors have many causes: touching characters, fragmented characters, the existence of non-character patterns, complicated arrangements, and so on. Obviously, an OCR-based method depends on a recognition process and suffers from low recognition accuracy for low quality images. However, OCR systems are still not available for the Lanna language. So, the OCR-based method is not a currently viable solution for processing Lanna document images.

The image-based method searches the document by comparing similarity between the keyword image and every word image in the document database. This generally consists of a pre-processing step and a searching step.

In the pre-processing step, a document image is segmented into word images and which are stored in a database along with their features. Next, the similarity between the keyword image and the word image is calculated by comparing their features in the searching step (Doermann, 1998). Some image-based searching approaches on the document images have been reported in the past few years. Lu *et al.* and Tan. (2002) represents each word as a string of feature codes and therefore each document image can be represented as a series of strings. An inexact string matching technique is used for matching a query word with the words in a document. A method described by Kim *et al.* (2005) uses character segmentation, feature extraction for the query keyword, and word-to-word matching to search for a keyword in Korean document images. Meshesha and Jawahar (2008) present a word image matching scheme for content-based document image retrieval from a printed text collection. Yadav and Sawarkar (2009) proposed a word image matching for searching user specified words in a document image by matching partial words. Each word image is represented by a primitive string and an inexact string matching technique is utilized to measure the similarity between the two 2 primitive strings.

The retrieval of historical printed documents has been developed in several years ago. Konidaris *et al.* (2007) proposed a technique for word spotting in machine-printed historical documents. The aim is to search a keyword in a large collection of digitized historical printed documents by matching synthetic keyword images with word images segmented from the processed document collections. The retrieval result is optimized by user feedback. A method for retrieval of Ottoman documents based on word matching is reported by Ataer and Duygulu (2006). A hierarchical matching technique is designed to find similar instances of the word images. The matching method consecutively tests length similarity and the similarity of quantized vertical projection profiles for entire words. However, although document image retrieval is a well-studied area, but this method has not yet extensively been applied extensively to study Lanna historical document images.

Unfortunately, the corpus and dictionary are still not available for Lanna OCR. So, an OCR engine is not a currently solution for Lanna word retrieval. Furthermore, the Lanna language has a complex writing system and we still lack of the expertise in the Lanna language. The Massive amounts of information from the Lanna document images are unreachable for those who are interested in them. Therefore, we proposed the a method for word retrieval form from old Lanna machine-printed document images. This research is to enable those who are interested in Lanna documents to accurately and easily browse or find the relevant information in a massive Lanna document image database.

In this paper, we present a method for word retrieval form old Lanna machine-printed document images. The proposed method involves a word matching process between synthetic keyword images with document word images. The aim is to find and locate this the keyword in the document image. The kKnowledge of the Lanna writing system such as the five 5 zones of a character is applied for synthetic keyword image creation. The retrieval effectiveness measures such as precision, recall, and F-measure are used for evaluating the performance of the

proposed method. Figure 1 illustrates a block-diagram for the proposed method based on: creating a synthetic keyword image, image preprocessing, and a word matching process.

The rest of the paper is organized as follows. First, we give a brief introduction on the old Lanna machine-printed documents images. Then we describe a method of creating synthetic keyword images, image preprocessing, and the word matching process. Finally, we discussing the experimental results and conclude the paper.

Materials and Methods

Old Lanna machineMachine-printed Printed document Document imagesImages

Digital images of old Lanna machine-printed documents have several different types of defects including darkened paper, adhesive marks, back-to-front interference, missing text, faded ink, dark borders, and textual noise, as shown in Figure 2(a). Furthermore, a large number of degraded texts have broken characters, non-uniformly spaced characters, and touching characters, as shown in Figure 2(b). It These makes this an even more a challenging task for

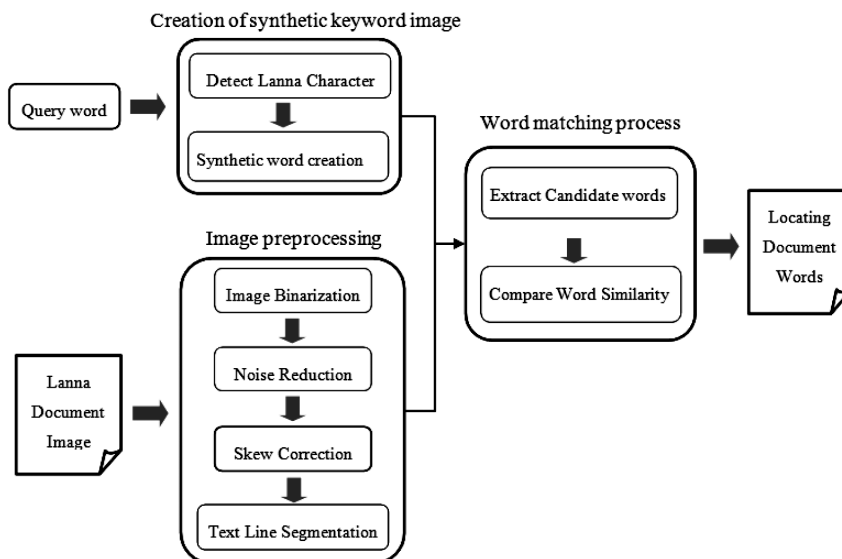


Figure 1. A block-diagram for word retrieval from old Lanna machine-printed document images

Lanna word searching.

In this research, we selected old Lanna machine-printed document images from a large amount of existing paper documents that were transferred to digital document images using a scanner (Chaijaruwanich, 2009). These documents are assumed to contain only text and are thus free from graphics, figures, maps, and tables. Figure 2(c) illustrates an example of the old Lanna machine-printed documents used in our experiments.

Creation of synthetic Synthetic keyword Keyword imagesImages

Synthetic image creation concerns the synthesis of a keyword image from the a Lanna word which is typed in the LN TILOK font typing format (Saengbun, 2016). Next, map their ASCII keyword is mapped to individual character image templates, then create a word image is created based on five the 5 zones of the character, the spacing between the characters, and the character arrangement. Figure 3 illustrates

an example of the resulting synthetic keyword image and their synthetic condition. A Lanna word structure is consists of five 5 zones. A character may occupy in one of the five 5 zones, namely: the upper zone 1 (UZ1), upper zone 2 (UZ2), central zone (CZ), lower zone 1 (LZ1), and lower zone 2 (LZ2), as shown in Figure 3(b). The spacing between the characters experimentally is defined as the x-value and y-value that has have been estimated over the document collection, as shown in Figure 3(c). This spacing value may vary depending on the document collection under study. The character arrangement uses the fitting characters' position for to integrate a single character image template into a synthetic keyword image. The fitting value is estimated by examine examining the character zone and other corresponding characters, as shown in Figure 3(d).

Image preprocessingPreprocessing

The old Lanna machine-printed documents are gray scale images that are characterized by

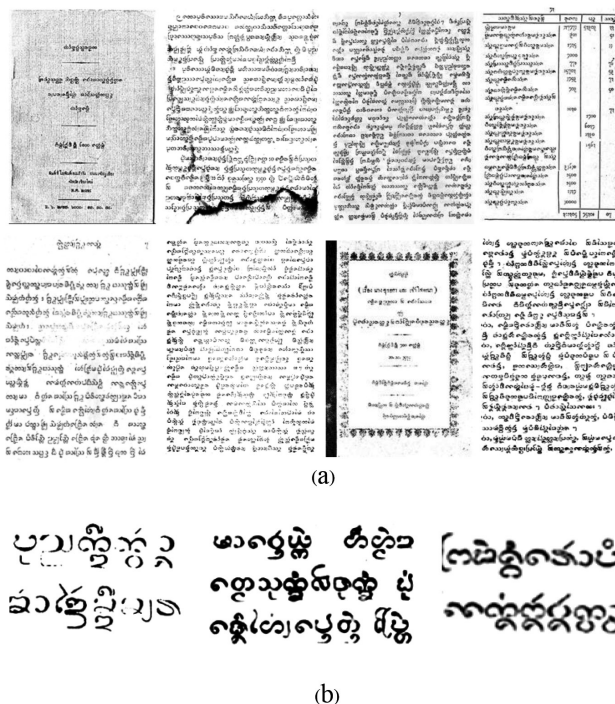


Figure 2. An example of old Lanna machine-printed document images: (a) some document defects, (b) degraded texts, and (c) a document image used for analysis

low image quality. Therefore, a preprocessing procedure is essential in order to improve the quality of these documents. The preprocessing process consists of four 4 distinct steps.

Firstly, image binarization is applied to the documents. This involves the conversion of the gray scale images into binary images. A proper threshold for the document image is selected by the Otsu's thresholding method (Otsu, 1979) and then converts all the intensity values that are above or below the threshold intensity are converted to one 1 intensity value representing either a "black" or "white" value.

The second step, the connected component labeling process (Haralick and Shapiro, 1992) process is used for reducing noise in the document image. This technique is proposed by firstly detecting the region of every connected component in the documents. The obtained connected components whose area is smaller than the defined threshold will be replaced by the white pixels.

The third step of the preprocessing process is the skew correction. It is necessary to identify and correct the text orientation and skew before proceeding to the text line segmentation step.

The final step of the preprocessing process is the text line segmentation process. The process uses two 2 complementary methods, projection profiles and the line separation method. At As a first step, we calculate the horizontal projection profile by summing the foreground pixels in every scan line. Then, the text lines' boundaries

are defined by calculate calculating the local minima. Next, the horizontal length threshold is experimentally defined as 150% of the average text line-height that has been estimated over the documents' collection. The text lines' boundaries that are greater than the horizontal length threshold are separated by using the line separation method.

Word matching Matching process Process

The word matching process consists of 2 distinct steps, candidate words extraction and word similarity comparison. In the candidate words extraction step, we propose 3 methods to determine the candidate words from the document image as follows.:

1. The pixel-based sliding window template matching (PSTM) method: following the sliding window approach, a synthetic keyword image is used as a template. A window template traverses throughout the text line from left to right and top to bottom. At each position of the window, a pixel density is computed using a bitwise AND operator. The function of the template matching process is defined as follows:

$$Match(w) = \frac{D_{img}}{D_{query}} \times 100 \quad (1)$$

$$D_{img} = \sum_{i=1}^m \sum_{j=1}^n AND \{q(i, j), w(i, j)\} \quad (2)$$

$$D_{query} = \sum_{i=1}^m \sum_{j=1}^n q(i, j) \quad (3)$$

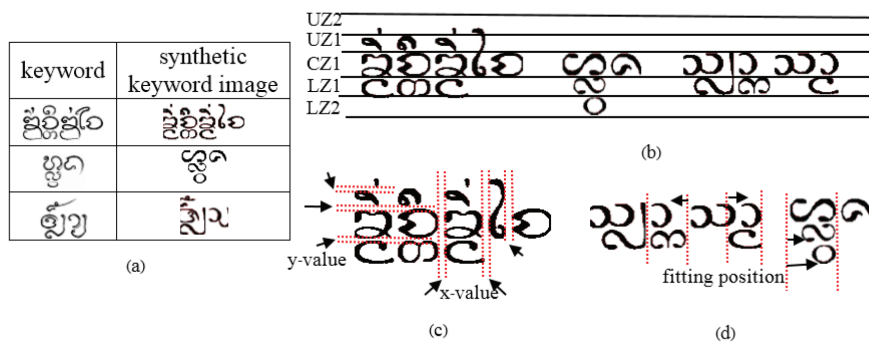


Figure 3. An example of synthetic keyword images: a) The creation of the synthetic image, b) five 5 zones of characters, c) spacing between the characters, and d) fitting characters' positions

where D_{img} is the similarity between a synthetic keyword image (q) with a given document word image (w) which is computed by using a bit wise AND operator over the entire image, m represents the width and n represents the height of the window, and D_{query} is the pixel density which is defined as the summation of black pixels in the synthetic keyword image. If $Match(w) > 50$, the document word image is regarded as a candidate word.

2. The block-based sliding window template matching (BSTM) method: Firstly, we produce the character blocks of a text line document image using the vertical projection method. Figure 4(a) shows the histogram which is obtained by counting the number of black pixels in each vertical scan and the columns where zero black pixels are used as delimiters for character blocks' separation. The character block images are shown in Figure 4(b). Next, the synthetic keyword image is used as a template to traverse the text line from left to right by sliding block-to-block, then the same functions as described in the PSTM method (Equation 1-

Equation 3) are used to evaluate the candidate word images.

3. The word feature matching (WFM) method: we use synthetic keyword features to evaluate the candidate words by comparing their features. After the character block separation, the features of a synthetic keyword image are generated for the feature of each character in the word that depends on the character height, as shown in Figure 5(a) and width, as shown in Figure 5(b).

This synthetic keyword feature is defined as follows:

$$F_q = [H_q, W_q] \quad (4)$$

where H_q and W_q are, respectively, the estimated height and width of the synthetic keyword image.

H_q and W_q are defined as follows:

$$H_q = \max(upper) - \min(lower) \quad (5)$$

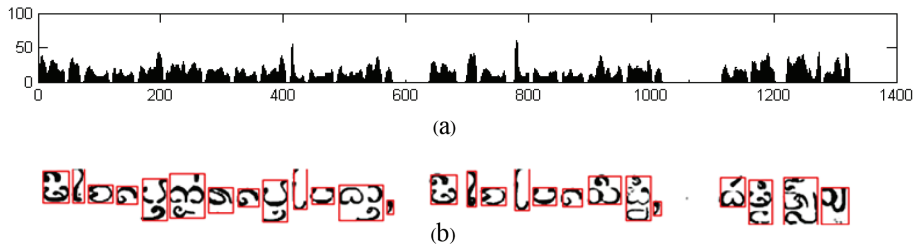


Figure 4. The separation of character blocks: (a) vertical projection histogram and (b) character block images

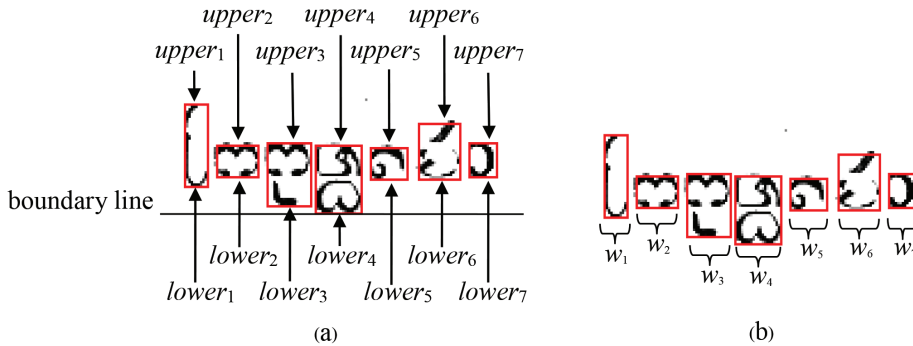


Figure 5. The features of a synthetic keyword image: (a) character height extracted with boundary line and (b) character width extracted with character blocks' image

where *upper* and *lower* are, respectively, the upper bound and lower bound of each character block.

$$W_q = \sum_{i=1}^n w_i \quad (6)$$

where *upper* is the width of the i^{th} character block and *n* is the total number of character blocks in the synthetic keyword image.

Based on the synthetic keyword feature extraction described above, we can define a synthetic keyword feature. Afterwards, we extract the candidate words from the document image by comparing the synthetic keyword feature with the character blocks feature from the document image. The function of the comparing process is defined as follows:

$$Match(F_q, b_{i \rightarrow j}) = \begin{cases} 1 & \text{if } (\max(upper(b_{i \rightarrow j})) - \min(lower(b_{i \rightarrow j}))) = H_q \\ & \text{and } \sum_{k=1}^l w_k = W_q \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

where $b_{i \rightarrow j}$ is the i^{th} to j^{th} character blocks from the document image, *upper* ($b_{i \rightarrow j}$) and *lower* ($b_{i \rightarrow j}$)

are, respectively, the upper bound and lower bound of each i^{th} to j^{th} character blocks, and w_k is the width of the k^{th} character block. If *Match* ($F_q, b_{i \rightarrow j}$) = 1, the i^{th} to j^{th} character blocks image is regarded as a candidate word.

After, we evaluate the candidate words from the document image using the methods described above. A word similarity comparison step is required. In our approach, we employ 2 methods to compare the similarity between synthetic keywords and word images. The first one is the Ssub-window similarity and the second is the Euclidean distance.

In the case of the Ssub-window similarity, the word image is divided into $m \times n$ uniform blocks (sub-window), where *m* and *n* are the width and height of the word image which are divided by 9. The sub-window density is calculated by counting the number of black pixels in each sub-window, and the density of the whole image is represented as an $m \times n$ matrix with each element standing for the density of the corresponding sub-window, and the matrix is called the density matrix. Let $W = [w_1, w_2, w_3 \dots w_{m \times n}]$ and $Q = [q_1, q_2, q_3 \dots q_{m \times n}]$ be a synthetic keyword

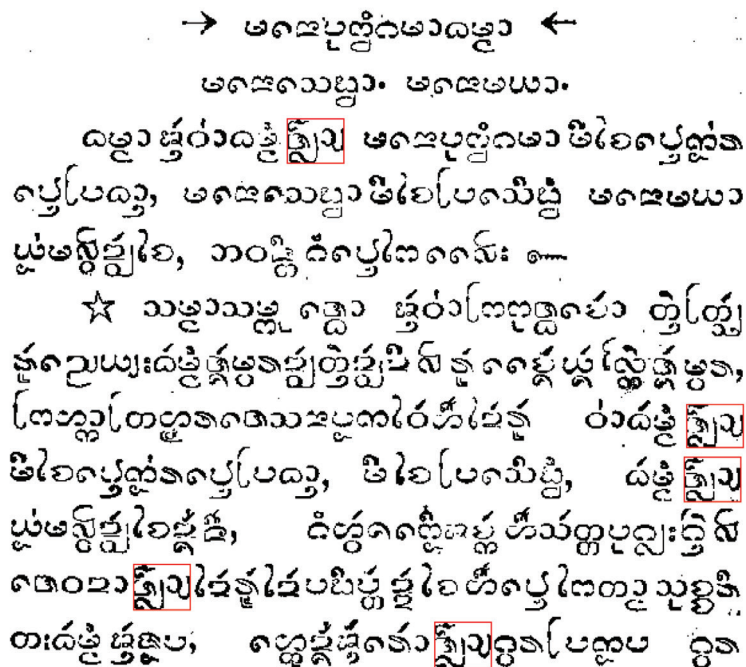


Figure 6. An example of the retrieval results for the word “နှိပ်”

and word image density matrix, respectively. The function of comparing each sub-window density in the density matrix is defined as follows:

$$Count(k) = \begin{cases} 1 & \text{if } (\frac{w_k}{q_k} \times 100) \geq 50 \text{ or } w_k = q_k = 0 \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

where w_k and q_k refer to the corresponding density in the k^{th} sub-window of the word images and the synthetic keyword image, respectively. If $\sum_{k=1}^{m \times n} Count(k) \geq T_w$, the word image is argued to be the desired one, where the threshold is determined experimentally.

In the case of the Euclidean distance, the word image is also divided into an $m \times n$ sub-window where m and n are the width and height of the word image and are divided by 16. The density of the whole image is represented as an $m \times n$ matrix with each element standing for the density of the corresponding sub-window. Let $W = [w_1, w_2, w_3 \dots w_{m \times n}]$ and $Q = [q_1, q_2, q_3 \dots q_{m \times n}]$ be a synthetic keyword and word image density matrix, respectively, where each sub-window density can be expressed as follows:

$$\text{sub-window density} = \frac{\text{the number of black pixels in sub-window}}{\text{the number of all pixels in sub-window}} \quad (9)$$

Next, the Euclidean distance is used to evaluate the similarity between a synthetic

keyword and the word image density matrix. The function for the Euclidean distance is defined as follows:

$$Edist(W, Q) = \|W - Q\|_2 = \left(\sum_{k=1}^{m \times n} (w_k - q_k)^2 \right)^{\frac{1}{2}} \quad (10)$$

where w_k and q_k refer to the corresponding density in the k^{th} sub-window of the word images and the synthetic keyword image, respectively. If $Edist(W, Q) \leq T_w$, the word image is argued to be the desired one, where the threshold T_w is determined experimentally.

Results and Discussion

The experiments involved a search for 197 keywords. An evaluation of the results was performed using precision, recall, and the F-measure. Precision is the ratio of the number of retrieved relevant words to the number of retrieved words. Recall is the ratio of the number of retrieved relevant words to the number of total relevant words marked on the sample document pages. The F-measure can be interpreted as a weighted average of the precision and recall. The precision, recall, and F-measure are defined as follows:

$$Precision(D) = \frac{R_{rw}}{R_{img}} \quad (11)$$

$$Recall(D) = \frac{R_{rw}}{R_{mw}} \quad (12)$$

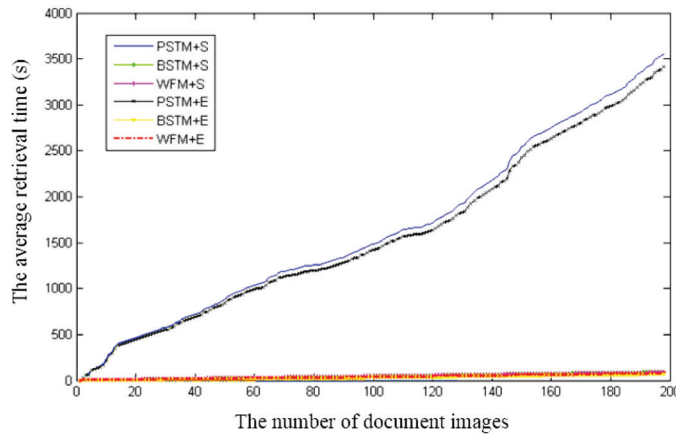



Figure 7. This graph compares the average time taken by six 6 different methods

Table 1. The average precision/recall rates and F-measures

Experiments	Average precision	Average recall	F-measure
PSTM+S	0.65	0.92	0.76
BSTM+S	0.73	0.81	0.77
WFM+S	0.73	0.74	0.73
PSTM+E	0.82	0.81	0.81
BSTM+E	0.80	0.78	0.79
WFM+E	0.84	0.80	0.82

$$F - measure = 2 \times \frac{precision \times recall}{precision + recall} \quad (13)$$

where R_{rw} denotes the retrieved relevant words from the document image D , R_{img} denotes the number of retrieved word images, and R_{rw} denotes the total number of relevant marked words. The *precision* and *recall* are the average of the precision and recall for all keywords, respectively.

Based on the words extraction from the document image by using the methodologies: Pixel-based Sliding Window Template Matching (PSTM), Block-based Sliding Window Template Matching (BSTM), and Word Feature Matching (WFM) methodologies and the comparison of word similarity by using the methods: Ssub-window similarity (S) and Euclidean distance (E) methods., We defined the experiments as follows: PSTM+S, BSTM+S, WFM+S, PSTM+E, BSTM+E, and WFM+E. These experiments are a combination of the candidate words extraction method and the word similarity comparison method, as described above. Figure 6 shows an example of the word retrieval result with the keyword “”. For each the retrieval method we consider the complexity required for computing the distances. To analyze the complexity of the proposed methods, we compare the computational cost with the average retrieval time, as shown in Figure 7. It is clearly seen that the PSTM+S and PSTM+E methods provide the inferior efficacy. We have also shown the qualitative results on six 6 different retrieval methods in Table 1. The performance of the retrieved results showed that the WFM+E method performs the best among the set of experiments.

Conclusions

This paper proposes a method for word retrieval from old Lanna machine-printed document images. The proposed method introduces the word retrieval mechanism through the creation of synthetic keyword images along with a hybrid method to enable a successful word matching process. This process requires the creation of a synthetic keyword image, image preprocessing, and then a word matching process in which a combination of candidate words' extraction and word similarity comparison methods are used. From our experimental results, the best performance was found for the method which was a combination of W word F feature M matching and the Euclidean distance method. The retrieval effectiveness measures confirm that the proposed method is feasible, valid, and effective for Lanna word image searching.

However, the present method is only appropriate for document images which only contain text and are free from graphics, figures, maps, and tables. Furthermore, the current method is only applicable if the document image has only a few defects and is not degraded. This assumption does not hold for all old Lanna machine-printed document images from the full historical collection.

References

- Ataer, E., and Duygulu, P. (2006). Retrieval of Ottoman documents. Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval; October 26 - 27, 2006; Santa Barbara, California, USA, p. 155-162.
- Chaijaruwanich, J. (2009). Wannapim Lanna. Chiang Mai University, Chiang Mai, Thailand, p. 5-185.

- Doermann, D. (1998). The indexing and retrieval of document images: a survey. *Comput. Vis. and Image Understanding*, 70(3): 287-298.
- Haralick, R. M., and Shapiro, L. G. (1992). *Computer and Robot Vision*. 1st ed. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 672p.
- Kim, S. H., Park, S. C., Jeong C. B., Kim, J. S., Park, H. R., and Lee, G. S. (2005). Keyword Spotting on Korean document images by matching the keyword image. *Proceedings of the 8th International Conference on Asian Digital Libraries (ICADL 2005)*; December 12-15, 2005; Bangkok, Thailand, p. 158-166.
- Konidaris, T., Gatos, B., Ntzios, K., Pratikakis, I., Theodoridis, S., and Perantonis S. J. (2007). Keyword-guided word spotting in historical printed documents using synthetic data and user feedback. *Int. J. on Doc. Anal. and Recognition (IJ DAR)*, 9(2): 167-177.
- Lu, Y., and Tan, C. L. (2002). Word searching in document images using word portion matching. *Proceedings of the 5th International Workshop on Document Analysis Systems*; August 19-21, 2002; Princeton, NJ, USA, p. 319-328.
- Maneechedta, W. (1994). *Enculturation process on Lanna dhama alphabets*, Master's [M.Ed. Thesis]. Department of Nonformal Education, Graduate School, Chiang Mai University, Chiang Mai, Thailand, 79p. (in Thai)
- Marukawa, K., Hu, T., Fujisawa, H., and Shima, Y. (1997). Document retrieval tolerating character recognition Errors-Evaluation and application. *Pattern Recognition*, 30(8): 1361-1371.
- Meshesha, M., and Jawahar, C. V. (2008). Matching word images for content-based retrieval from printed document images. *Int. J. of Doc. Anal. and Recognition (IJ DAR)*, 11(1): 29-38.
- Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE T. on Syst., Man and Cybernetics*, 9(1): 62-66.
- Pansook, A. (2008). *Status and role changes of the Lanna language*, Master's [M.A. thesis]. Department of Lanna Language and Literature, Graduate School, Chiang Mai University, Chiang Mai, Thailand, 187p. (in Thai)
- Prongthura, N. (1982). *Dramma scripts of Northern Thailand*, Master's [M.A. Thesis]. Department of Oriental Language, Graduate School, Silpakorn University, Bangkok, Thailand, 313p. (in Thai)
- Saengbun, P. (2016). *Font Lanna "LN TILOK"*. Available from: <http://art-culture.chiangmai.ac.th/fontlanna>. Accessed date: Sep 30, 2016.
- Senda, S., Minoh, M., and Ikeda, K. (1993). Document image retrieval system using character candidates generated by character recognition process. *Proceedings of the 2nd International Conference on Document Analysis and Recognition (ICDAR '93)*; October 20-22, 1993; Tsukuba, Japan, p. 541-546.
- Yadav, S., and Sawarkar, S. (2009). Retrieval of information in document image databases using partial word image matching technique. *Proceedings of the IEEE International Advance Computing Conference (IACC 2009)*; March 6-7, 2009; Patiala, India, p. 552-557.