2012 International Conference on Frontiers in Handwriting Recognition

Text Line Extraction Using Adaptive Partial Projection for Palm Leaf Manuscripts from Thailand

Rapeeporn Chamchong, Chun Che Fung
School of Information Technology
Murdoch University
Murdoch, Western Australia, Australia
rapeeporn.c@gmail.com, l.fung@murdoch.edu.au

***Abstract*— Text line extraction is one of the critical steps in document analysis and optical character recognition (OCR) systems. The purpose of this study is to address the problem of text line extraction of ancient Thai manuscripts written on palm leaves, using an Adaptive Partial Projection (APP) technique by integrating a modified partial projection and smooth histogram with recursion. The proposed approach was compared with a Modified Partial Projection (MPP) looking at vowel analysis and touching components of two consecutive lines. The results from this research suggested that the proposed approach for practical data on palm leaf manuscripts has better performance in solving the line segmentation problem.**

Keywords-line extraction; line segmentation; historical document; document analysis system

I. Introduction

In the processing of ancient hand written manuscripts from Thailand, text line extraction is required to separate text lines and individual characters in the document. This is a significant step in pre-processing for document analysis and character recognition. Text line segmentation should then be followed by word segmentation and character segmentation. In the character recognition process, flow of the text components such as words, characters or alphabets may not be read correctly unless they are in proper sequence. Consequently, line extraction is needed to form a horizontal script, enabling it to be recognized or read properly.

In order to process a large volume of ancient manuscripts which contain valuable knowledge and information, there is a need for automated systems that are capable to work with practical documents in an efficient and accurate manner. For example, dried palm leaves had been used as one of the most popular writing media in Thailand during the past centuries. Such documents are heritage from past civilization passed down through many generations. At present, there is no specific system that can process practical handwritten document of Thai language because it is different from other language systems. The use of specific tonal, vowel and consonant characters with multiple levels and the lack of word spacing are the key challenges in the automatic processing of Thai language documents. It is therefore the main objective of this study to develop an efficient and intelligent image processing system that could be used to extract components from these ancient manuscripts for information retrieval and preservation purposes. In this study, the language on the palm leaf manuscripts acquired from the Project for Palm Leaf Preservation in Northeastern Thailand Division, Mahasarakham University [1] are Thai-Noi, which is different from the modern Thai language.

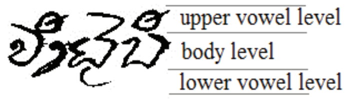


Figure 1. Three levels of Thai-Noi writing.

At present, there is not much reported work on the handling of horizontally overlapping lines in modern Thai and ancient Thai handwriting. In this study, Thai-Noi written on palm leaf manuscripts is comprised of three levels, which are the upper vowel, body, and lower vowel level as shows in Fig. 1. Due to the three levels required to form a text line, this affects the line separation process. In general, Thai-Noi writing starts from left to right and from top to bottom. It does not require spaces between words and sentences.

In the survey by Likforman-Sulem et al. [2], text line segmentation of historical documents is separated into six categories, they are: projection-based, smearing, grouping, Hough-based, repulsive-attractive network and stochastic methods. They reported that piecewise projections proposed by Pal and Datta [3], and Zahour et al. [4] are suitable for overlapping or touching lines, and another technique based

978-0-7695-4774-9/12 $26.00 © 2012 IEEE
DOI 10.1109/ICFHR.2012.280

on stochastic method proposed by Tseng and Lee [5], is also suitable for overlapping lines and it is more robust. Their summary stated that there is no single line segmentation technique that suits all historical documents. The particular technique will depend on the characteristics of the writings such as script size, stroke width and average spacing.

Surinta [6] proposed sorting and distinguishing on the basis of projection profile. This was experimented with single column of Thai handwritten documents. The accuracy of this technique achieved was 97.11%. However, this experiment did not consider overlapping consecutive lines and fluctuating lines.

Arivazhagan et. al [7] proposed a statistical approach to line segmentation. Their results showed that on 720 documents (English, Arabic, and children's handwriting) containing 11,581 lines, the approach segmented correctly at 97.31% and there were over 200 handwritten images with 78,902 connected components, 98.81% of them were associated with the correct lines. Most of the errors were due to two reasons: normal component, which spans across two or more lines, and normal component, lying in between two lines. Their technique could preserve the dot above and below a word.

This study is aimed at text line extraction on palm leaf manuscripts using an improved technique by enhancing partial projection profile through the integration of a modified projection and smooth histogram with recursion. In this technique, width size of partial projection and height size of character for smooth histogram are adapted by calculation based on all the individual characters in each image. During the line processing, if each partial projection cannot separate the lines properly, partial projection will be divided into two columns and execute recursively in each column. The next section explains the MPP method and the proposed APP technique. Experimental results and discussion are presented in Section III, and then followed by the conclusion and consideration for future work.

II. LINE EXTRACTIONS

In this study, two line extraction techniques were compared. A modified partial projection looking at vowel analysis and touching components of two consecutive lines [8] is explained in section A, and an improved technique by adapting the modified partial projection and smooth the histogram with recursion is described in section B.

A. Modified Partial Projection (MPP) Technique

To separate text lines, the partial projection technique [4, 9] is applied by dividing the text images into vertical columns. The MPP approach [8] to separate the lines is outlined as follows:

1. Divide the image into vertical columns by using the average width of the characters which is calculated from the mean value of the width of the characters in the data set from the palm leaf manuscript images.
2. Find the horizontal projection profile ($P[y], y \in \{1,2,3,...,row\}$) along the horizontal axis for each row of y in each column.

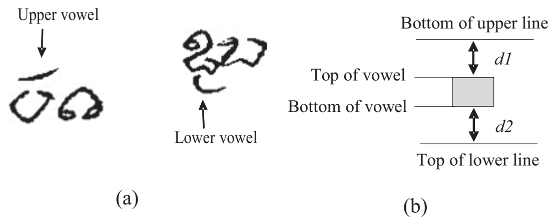


Figure 2. (a) Consonant with vowel, (b) analyzing the distance of the vowels.

3. Find the minimum value of the projection in each column. This minimal value of histogram indicates the top and bottom lines. A bottom line is chosen as a base line. If the height in each line is less than the average character height (*Ht*), which is calculated from the mean value of character heights from the data set, this based line is deleted. However, there are some vowels appear above or below the characters and they were drawn as isolated components as shown in Fig. 2 (a). The positions of these vowels occupy certain distance from the characters. This significantly affects the separating line. To calculate this value, two distances are calculated as shown in Fig. 2 (b). The value of *d1* defines the distance between the bottom of the upper line and the top of the vowel. *d2* defines the distance between the bottom of the vowel and top of the lower line. If ($d1 \geq d2$) then this vowel belongs to the lower line. If ($d2 > d1$) then this vowel belongs to the upper line.

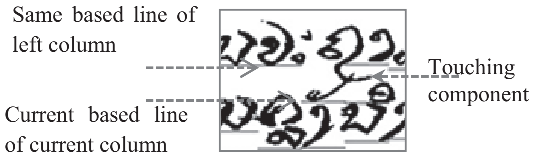

Same based line of left column
Touching component
Current based line of current column

Figure 3. An example of touching components between two consecutive lines.

4. Calculate the average value of a number of lines (*avg_num_line*) of all columns and check the number of lines in each column. If the number of lines in each column is less than the *avg_num_line*, a base line is added from the same line of the closest right/left column which has a lower base line. This process starts from the right column to the left column. If the components of two lines are connected, they can be separated by checking against the gap between the two lines as shown in Fig. 3. Touching consecutive lines can be separated into two lines by setting the base line position of the upper line as in Equation (1).

$$line[i][j] = line[k][j] + \frac{|line[i][j]\ \text{-}\ line[k][j]|}{2} \tag{1}$$

where i is the current column, j is the current line position, and k is the left line position.

5. Join the horizontal line and then form a separate line.

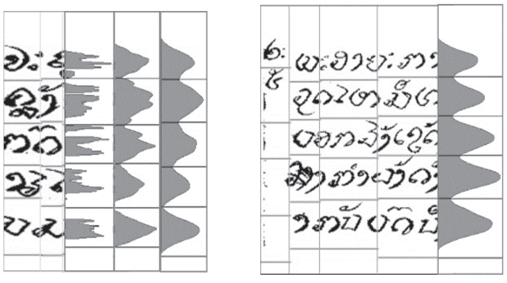
B. Adaptive Partial Projection (APP) Technique

The proposed technique is derived from the partial projection method [4, 9]. The text image is divided into vertical columns and then the histogram of the projection profile is applied by smoothing [7] to separate the lines. Smoothing is used to remove spurious peaks and valleys of the histogram. A moving average filter is applied for smoothing based on the height of the characters. In this technique, is calculated from the average value of the height of the characters in each palm leaf manuscript image. The characters in each image were extracted by using Connected Component and only individual components were selected automatically. The size of a partial column uses three characters which are the common length for a word in Thai-Noi script. The width of a character is calculated from the average value of the width of the characters in each image. However, if the line cannot be separated, recursion will then be applied to divide the column into two and the process is iterated to find the base line again until the column size is less than 75% of the width of the character. This can reduce the time for the line extraction process because this method is based on three characters instead of one. The approach to separate the lines by this technique is described as follows:

1. Find the number of lines (L), the average line position of each line (μ_k), and the average height of lines (Ht_L) from the global horizontal projection:
   - Calculate the horizontal projection profile of the image.
   - Smooth the histogram by moving average filtering with Ht .
   - Find the peaks of the histogram and then define L as the number of lines from the number of peaks.
   - Then define the average line position (μ_k) of each line as shown in Equation 2.

$$\mu_k\,,\ k \in \{1, 2, \ldots, L\} \tag{2}$$

   - Calculate Ht_L by averaging the different values between μ_k
2. Divide the image into vertical column. The width of the column is defined as three characters which were calculated automatically in each image.
3. Calculate the horizontal projection profile ($P[y], y \in \{1, 2, \ldots, rows\}$) along the horizontal axis for y values in each column as shown in Fig. 4 (a).
4. Smoothen the histogram ($SP[y]$) twice by moving average filtering with Ht which were calculated automatically in each image as shown in Fig. 4 (a). The moving average filtering is applied twice because the spurious peaks and valleys in the projection occurred after the first smooth of the histogram.
5. Find the base lines ($\beta_i, i \in \{1, 2, \ldots, N\}$) in each column by using the following rules:



(a) (b)

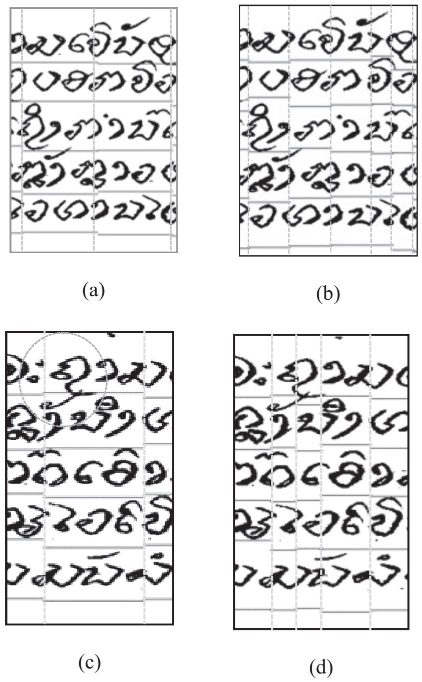
Figure 4. Samples of smooth histogram and base lines in each column (a) Projection, the first smooth histogram and the second smooth histogram (left to right) of the second column
(b) The final smooth histogram of the last column

   - Find the valleys of the smooth histogram as shown in Fig. 4 (b). A valley of histogram is the lowest point between two peaks. These valleys are defined as base lines in each column. In Fig. 4 (b), four valley values are shown and they are used together with the last lowest value at the bottom of the diagram.
   - Test all the valleys. If two consecutive valleys are very close (length between β_i and β_{i+1} less than $(4/5)\,Ht$), this can be assumed that it is a base line of vowel. To select the candidate valleys, projection profile ($P[y]$) is used to test as follows:
     (1) If $P[\beta_i] > 0$ and $P[\beta_{i+1}] > 0$ then go to (2), otherwise go to (3)
     (2) If $SP[\beta_i] < SP[\beta_{i+1}]$ then delete β_i , otherwise delete β_{i+1}
     (3) If $P[\beta_i] = 0$ and $P[\beta_{i+1}] = 0$ then go to (4), otherwise go to (5)
     (4) If $(\beta_i - \beta_{i-1}) < Ht$ then delete β_i , else if $(\beta_{i+2} - \beta_{i+1}) < Ht$ then delete β_{i+1} , otherwise set $\beta_i = \beta_i + (\beta_{i+1} - \beta_i)/2$ and delete β_{i+1}
     (5) If $P[\beta_i] = 0$ then delete β_{i+1} , otherwise delete β_i
   - Check for incorrect top line (this may be upper vowel of the first or unnecessary information) and bottom line (this may be lower vowel of the last line or unnecessary information) as follows:
     (1) Examine the top line if there is $\beta_i > \mu_1$ then delete β_i .
     (2) Examine the bottom line if there is $\beta_{i+1} > \mu_k$ then delete β_{i+1} .

- Test the number of base lines (*N-1*) because palm leaf manuscripts may have some holes among a line as shown in samples in Fig. 6 and the gap at left or right borders of the image. Another reason is due to some base lines may occur because of vowels on the top and bottom of the characters. These need to be checked, and insert or delete the correct base line to each column. It is done as follows:

(1) If (*N-1*)<*L*, then a base line will be inserted by checking against μ_k that a base line (β_i) belongs to μ_k (i.e. β_1 belongs to μ_1 , β_2 belongs to μ_2 , ….). If there is no base line belonging to μ_k , then a base line at this position will be inserted by $\beta_i = \beta_i + Ht_L$

(2) If (*N-1*)>*L*, then a base line will be deleted by checking against μ_k . If there are more than a base line belonging to μ_k , then base line between mid of μ_k and μ_{k+1} will be deleted.



(a) (b)

(c) (d)

Figure 5. Sample of recursion (a) vowels and prolong problems (b) recursion when vowels and prolong occur (c) connected component and (d) recursion when connected component occurs

- Test for connected component, upper/lower vowel levels, or prolonged parts of consonants at all base line position (shows as Fig. 5) as follows:

(1) If $P[\beta_i] > 0$ then traverse the projection up and down from β_i position to the size of the vowel, *Hv* (This is estimated at half the height of character, *Ht/2*). Go to step (2).

(2) If the first position (F) of $P[\beta_i] = 0$ is between up traverse and down traverse and $|F - \beta_i| > \frac{1}{2}Hv$ then test the height size (H) between β_{i-1} and F . If $\frac{1}{2}H_L < H < 1\frac{1}{2}H_L$ then set the new position to β_i . A sample result is shown in Fig.5 (b).

(3) If $P[\beta_i] \neq 0$ then it is a connected component, go to step 6. A sample result is shown in Fig. 5 (d).

(4) If a base line overlaps one or more connected components, then divide the column into two and re-apply step 3 to step 6 to each of the column. The recursion is stopped when the width of the column is less than 75% of the width of a character.

(5) Join the horizontal line and then form a separate line.

III. Experimental Results and Discussion

The experimental results based on two line extraction techniques are presented in this section, (a) MPP and (b) APP. Palm leaf manuscript images were collected from the Project for Palm Leaf Preservation in Northeastern Thailand Division, Mahasarakham University [1]. The proposed system has been implemented with Visual C++ and OpenCV library. The resolution of the input images is 200x200 dpi in RGB format. The input images were transformed to binary image by automatically selecting the optimal binarisation algorithm by [10].

In the experiment, 264 text lines from 60 palm leaf manuscripts were considered. To check whether a text line is extracted correctly, a boundary is drawn between two lines. The result of line extraction is measured by following the rules in [3]. In Fig. 6, example results are shown and results of the experiments are given in Table 1. The experiment shows that the MPP and the proposed technique correctly segmented 129 lines of 264 lines (48.86%) and 177 lines (67.05%) respectively. The 69 lines of all lines (26.14%) and 57 lines of all lines (21.59%) respectively have one component out of the correct lines. For the lines with two components out of the correct lines, the number is 21 (7.95%) and 17 (6.44%) respectively. The rest are more than two out of the correct lines.

The APP technique can be used to separate some touching characters from consecutive lines. The APP technique integrates the MPP and smooth histogram with recursion so that the proposed method can cope with vowels as MPP technique. The proposed method also varies the window size of partial column and window size of height of character according to the image. This method adjusts the size of the character automatically in each image. The APP also reduces the time for the line extraction process because the method is based on three characters instead of one in the MPP, and the APP considers re-execution in small partial columns as required. This technique also checks the prolonged characters during the process. However, errors may be due to overlapping by the prolonged characters. Furthermore, some binary images are unclear and they have a major effect on the accuracy of text line extraction.

IV. Conclusion and Future Work

Line extraction is still one of the most challenge topics in document image analysis. In this paper, a proposed adaptive partial projection method has been compared with a modified partial projection method. The APP method provided better performance than MPP and it can be applied as a preliminary stage of a fully automated document analysis, retrieval and recognition system in the future. However, there are a few problems caused by (1) vowel which are too close to the upper or lower line than its own line, (2) long prolonged components, (3) a few connected components of consecutive lines are not separated into proper positions. As there are limited reports on the research of line extraction for Thai and ancient Thai manuscripts, this research will assist the development of an automated system which is in progress right now. Another challenge is there is no developed technique which has proven to have a high level of accuracy with practical data on ancient manuscripts written in Thai language. In future, improving the performance of the proposed technique could be achieved by considering the identification of prolonged components and connected components of consecutive lines. In order to verify the accuracy of the proposed system for the processing of ancient Thai manuscripts, more data sets will be used to test the prototype.

Acknowledgment

The authors wish to express their thanks and appreciation to members of the Preservation of Palm Leaf Manuscripts Project, Mahasarakham University, Thailand for their support and providing images from their database.

References

[1] Mahasarakham University, Project for Palm Leaf Preservation in Northeastern Thailand Division, Mahasarakham University, "Palm Leaf Manuscripts in Northeastern Thailand," Project for Palm Leaf Preservation in Northeastern Thailand, [Online]. Available: http://www.bl.msu.ac.th/2553/english_bl.htm. [Accessed: 27 February,2011].

[2] L. Likforman-Sulem, et al., "Text line segmentation of historical documents: a survey," International Journal Document Analysis and Recognition vol. 9, pp. 123 - 138, April 2007.

[3] U. Pal and S. Datta, "Segmentation of Bangla unconstrained handwritten text," in Proceedings of the Seventh International Conference on Document Analysis and Recognition, 2003, pp. 1128-1132.

[4] A. Zahour, et al., "Arabic hand-written text-line extraction," in Proceedings. Sixth International Conference on Document Analysis and Recognition, 2001, pp. 281-285.

[5] Y. H. Tseng and H. J. Lee, "Recognition-based handwritten Chinese character segmentation using a probabilistic Viterbi algorithm," Pattern Recognition Letters, vol. 20, pp. 791-806, 1999.

[6] O. Surinta, "Optimization of line segmentation techniques for Thai handwritten documents," in Eighth International Symposium on Natural Language Processing, 2009. SNLP '09. , 2009, pp. 180-183.

[7] M. Arivazhagan, et al., "A statistical approach to line segmentation in handwritten documents," in Proc. SPIE on Document Recognition and Retrieval XIV, CA, USA, 2007.

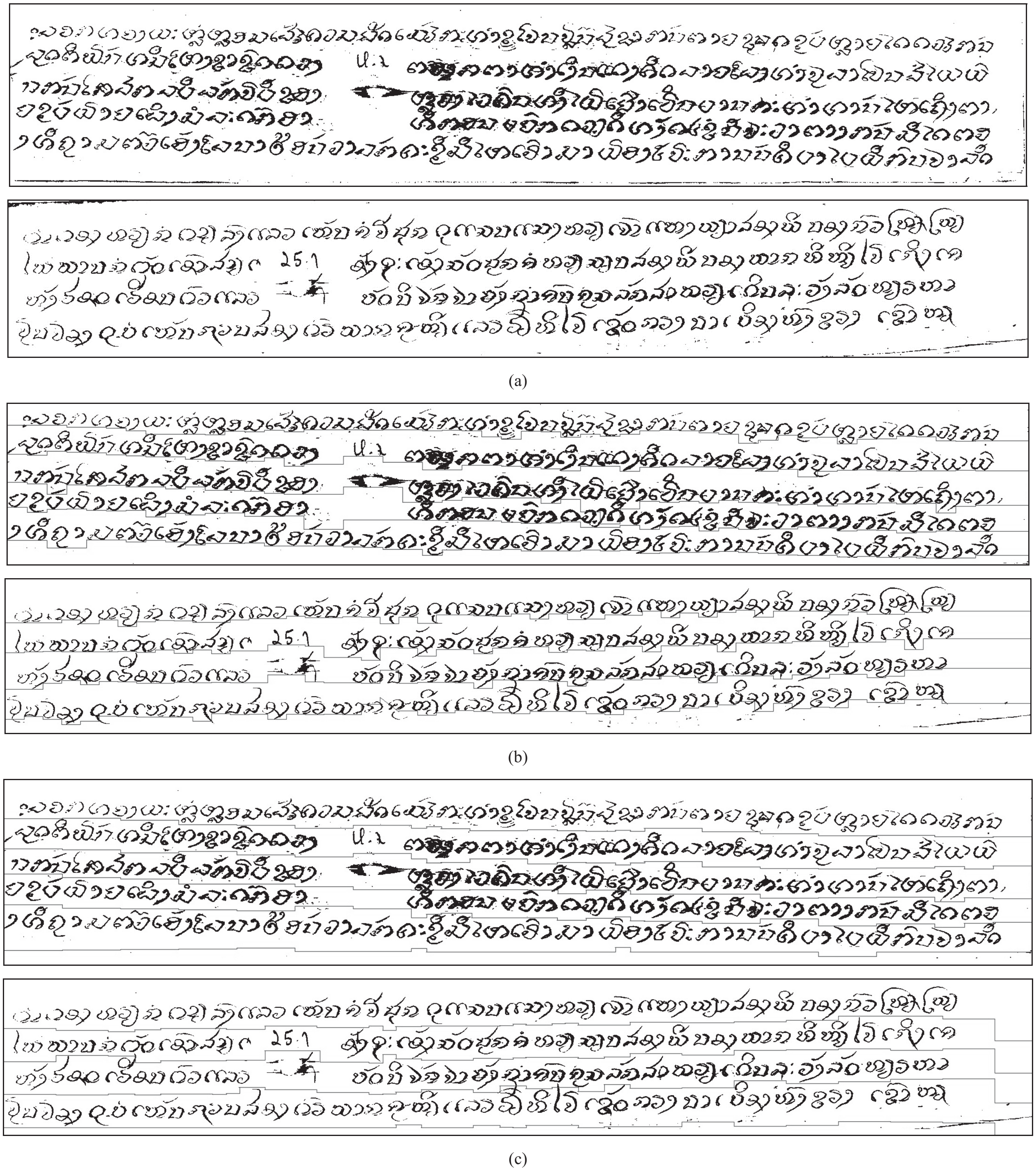
[8] R. Chamchong and C. C. Fung, "Character segmentation from ancient palm leaf manuscripts in Thailand," in Proceedings of the 2011 Workshop on Historical Document Imaging and Processing, Beijing, China, 2011.

[9] N. Tripathy and U. Pal, "Handwriting Segmentation of Unconstrained Oriya Text," presented at the Ninth International Workshop on Frontiers in Handwriting Recognition (IWFHR'04), 2005.

[10] R. Chamchong and C. C. Fung, "Optimal selection of binarization techniques for the processing of ancient palm leaf manuscripts," in 2010 IEEE International Conference on Systems Man and Cybernetics (SMC), Istanbul,Turkey, 2010, pp. 3796-3800.

TABLE I. RESULTS FROM TEXT LINE EXTRACTION

Number of components out from their correct line	**Percentage of components segmented within the correct line**	**MPP**		**APP**	
		Number of correct lines is done	***Percentage of correct line is done***	***Number of correct lines is done***	***Percentage of correct line is done***
0	100%	129	48.86%	177	67.05%
1	98.00-99.99%	69	26.14%	57	21.59%
2	96.00-97.99%	21	7.95%	17	6.44%
3	94.00-95.99%	22	8.33%	9	3.41%
≥4	<=95.99%	23	8.71%	4	1.52%



(a)

(b)

(c)

Figure 6. Sample results of line extraction (a) original images (b) MPP technique (c) APP technique