Text Line Segmentation of Ancient Thai Manuscripts on Palm Leaves

R. Chamchong¹² and C. C. Fung¹
¹School of Information Technology, Murdoch University, Perth, Western Australia, Australia
²Faculty of Informatics, Mahasarakham University, Maha Sarakham, Thailand

Abstract - Segmentation of handwritten scripts with overlapping text is one of the challenging tasks in the preprocessing for document recognition and optical character recognition (OCR) systems. It is a significant step because errors in the recognition stage will occur if text lines are not separated accurately. This paper aims to address the problem of text line segmentation of ancient Thai manuscripts written on palm leaves, in particular dealing with the issue of overlapping characters. The proposed technique is based on an integration of a partial projection method and smooth horizontal histogram with recurrence in each column. The performance evaluation of the proposed technique was compared with a modified partial projection profile. The experimental results from this study show that the accuracy of the proposed technique has a better performance. This technique will help to resolve the problem of text line segmentation for ancient Thai manuscripts on palm leaves.

Keywords: Document analysis, Ancient manuscripts, Historical documents, Text line segmentation

1 Introduction

Text line segmentation is a critical step in preprocessing for document recognition and character recognition systems because errors in the recognition stage will occur as a consequence if the text lines were not separated accurately. In the processing of ancient handwritten manuscripts, text line segmentation is needed to separate text lines and isolate the characters in the document. The step follow the text line segmentation will then be word segmentation and character segmentation. In the OCR process, the flow of text components, that is characters or alphabets, cannot be read properly unless they are ordered in proper sequence. Consequently, text line segmentation is essential for the formation of a horizontal script.

Prior literature suggests several approaches for text line segmentation. In the review by Likforman-Sulem et al. on text line segmentation of historical documents, it described several methods for separating printed or handwritten documents, broken and touching characters and a comparison of segmentation results. 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 [1].

Zahour et al.[2] proposed the partial projection profile. This method works well for the text with overlapping lines, incomplete lines and change in text orientation but the method does not deal with touching line. While an improved version of partial projection profile method has been explained by Pal and Datta [3] and Tripathy and Pal [4], for handwritten Bangla scripts, false text line separation occurred due to several neighboring text lines are connecting.

Surinta [5] proposed sorting and distinguishing the histogram of projection profile in order to select the base line. This was experimented with single column of Thai handwritten documents and the accuracy achieved was 97.11%. However, this experiment did not consider overlapping consecutive lines and fluctuating lines.

Arivazhagan et. al [6] proposed a statistical approach to line segmentation. Their experimental results showed that on 720 documents in English, Arabic, and children’s handwriting which contained 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. It was reported that their technique could preserve the dot above and below a word.

In order to process vast amount of valuable ancient manuscripts, there is a need for automated systems that are capable to work with practical documents in an efficient and accurate manner. In Thailand during the past centuries, dried palm leaves had been used as one of the most popular writing media. Currently, there is no specific system that can process practical handwritten document of Thai language because it is very 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, most of the alphabets on the palm leaf manuscripts acquired from the Project for Palm Leaf Preservation in
Northeastern Thailand Division, Mahasarakham University are Thai-Noi, which is different from the modern Thai language.

Figure 1. Three levels of Thai-Noi writing.

At present, there is limited reported work on the handling of horizontally overlapping lines in modern Thai and Ancient Thai handwriting. In this study, ancient Thai language (Thai-Noi) written on palm leaf manuscripts is different from other languages because a text line is comprised of three levels, which are the upper vowel, body, and lower vowel level as shown in Figure 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 and it does not require spaces between words and sentences.

This study is aimed at text line segmentation on palm leaf manuscripts using an applied technique by enhancing partial projection profile through the integration of projection method and smooth horizontal histogram with recursion. The next section explains the modified partial projection profile method and the proposed integrating partial projection profile technique. Experimental results and discussion are presented in Section 3, and then followed by the conclusion and consideration for future work.

2 Line Segmentations

In this study, two line segmentation methods were compared. The first method is a modified partial projection method looking at vowel analysis and touching components of two consecutive lines [8] and the second method is an applied technique that is improved by integrating the partial projection profile and smooth the histogram with recursion.

2.1 The Modified Partial Projection Method

To separate text lines, the partial projection method [2]-[4] is applied by dividing the text images into vertical columns and the modified 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. The width of the column is calculated from the mean value of the width of the characters in the data set from the palm leaf manuscript images.

2. Calculate the horizontal projection profile \( P_y, y \in \{1,2,3,\ldots, \text{row}\} \) along the horizontal axis for each row of \( y \) in each column.

3. Find the minimum value of the horizontal projection in each column as shown in Figure 2. 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 \( H_t \), 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 Figure 3(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 Figure 3(b). The value of \( d_1 \) defines the distance between the bottom of the upper line and the top of the vowel. \( d_2 \) defines the distance between the bottom of the vowel and top of the lower line. If \( d_1 \geq d_2 \) then this vowel belongs to the lower line. If \( d_2 > d_1 \) then this vowel belongs to the upper line.
4. Calculate the average value of a number of lines \((\text{avg}_\text{num}_\text{line})\) of all columns and check the number of lines in each column. If the number of lines in each column is less than the \(\text{avg}_\text{num}_\text{line}\), 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 Figure 4. Touching consecutive lines can be separated into two lines by setting the base line position of the upper line as in Equation (1).

\[
\text{line}[i][j] = \text{line}[k][j] + \frac{\text{line}[i][j] - \text{line}[k][j]}{2}
\]  

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

5. Join horizontal line by linking the positions of the horizontal line and draw the joined line and then form a separate line.

![Diagram of touching components between two consecutive lines](image)

**Figure 4.** An example of touching components between two consecutive lines.

### 2.2 The Proposed Technique

The proposed technique in this study is derived from the partial projection profile, by dividing the text images into vertical columns and then apply smoothing histogram of projection profile [6] to separate the lines. Smoothing is used to remove spurious peaks and valleys in the projection profile. A moving average filter is applied for smoothing based on the height of character \((H_t)\). The height of the character is calculated from the average value of the height of the characters in the data set from the palm leaf manuscript images. This value is similar to the one described in (1) of Section 2.1. The size of column used 5% of the width of the image which was proposed by Arivazhagan, et al. [6]. However, the window should be wide enough to include a minimum of 3 to 5 characters, which are common length for a word in Thai-Noi script. If the line cannot be separated, recursion will then be applied to divide into two columns and the process is iterated to find the base line again. This can reduce the time for the line segmentation process. The approach to separate the lines by this technique is outlined as follows:

1. Find the number of lines and average line position of each line from the global horizontal projection profile of the image which is based on the complete row of text.
   - Calculate the horizontal projection profile.
   - Smooth the histogram by moving average filtering with \(H_t\).
   - Find the peaks of the histogram and then set these positions as the average line position \((\mu_k)\) of each line as shown in Equation 2.

\[
\mu_k, k \in \{1, 2, ..., L\}
\]  

where \(L\) is the number of lines which calculate from the number of peak value.

2. Divide the image into vertical column. The width of the column is defined as 5% of the width of the image. The number of columns is defined as \(N\).

3. Calculate the horizontal projection profile \(P[y, y]_{y=1,2,..., \text{rows}}\) along the horizontal axis for each \(y\) values in each column as shown in Figure 5(a).

4. Smoothen the projection profile \(SP[y]\) 2 times by moving average filtering with \(H_t\) as shown in Figure 5(a). The moving average filtering is applied twice because the spurious peaks and valleys in the projection profile occurred after the first smooth of the histogram.

![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, and (b) the final smooth histogram of the last column](image)

**Figure 5.** 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, and (b) the final smooth histogram of the last column

5. Find the base lines \((\beta_i, i \in \{1,2,...,M\})\) in each column by using the following rules:
   - Find the valleys of the smooth projection as shown in Figure 5(b). A valley of projection profile is the lowest point between two peaks. These valleys are defined as
base lines in each column. The first valley will not be used to define the base line.

Test all the valleys. If two consecutive valleys are very close (length between $\beta_i$ and $\beta_{i+1}$ less than $\frac{3}{4}Ht$), 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 $\beta_i$, otherwise delete $\beta_{i+1}$.

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 $\beta_i$, else if $\beta_{i+2} - \beta_{i+1} < Ht$ then delete $\beta_{i+1}$, otherwise set $\beta_i = \beta_i + (\beta_{i+1} - \beta_i)/2$ and delete $\beta_{i+1}$.

5. If $P[\beta_i]=0$ then delete $\beta_{i+1}$, otherwise delete $\beta_i$.

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 $\beta_i > \mu_i$ then delete $\beta_i$.

2. Examine the bottom line if there is $\beta_{i+1} > \mu_k$ then delete $\beta_{i+1}$.

Test the number of base lines because palm leaf manuscripts may have some holes among a line as shown in samples in Figure 7 and the gap at left or right borders of the image so these need to check and insert the correct base line to each column. It is done as follows:

1. If the number of base lines ($M-1$) is less than the number of lines ($L$), then a base line will be inserted by checking with $\mu_i$ that a base line ($\beta_i$) belongs to $\mu_i$ ($\beta_i$ belongs to $\mu_i$, $\beta_j$ belongs to $\mu_2$, ...).

2. If there is no base line belonging to $\mu_k$, then a base line at this position will be inserted by $\beta_{i-1} + (\beta_i - \beta_{i-1})/2$.

Test for connected component, upper/lower vowel levels, or prolonged parts of consonants at all base line position (shows as Figure 6) as follows:

1. If $P[\beta_i]>0$ then traverse the projection up and down from $\beta_i$ 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 of $P[\beta_i]=0$ is between up traverse and down traverse then set new position to $\beta_i$. A sample result is shown in Figure 6(b).

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

6. If a base line overlaps one or more connected components, then divide the column into two and recursive itself from step 3 to step 6. The recursion is stopped when the width of the column is less than 1% of the width of the image or less than the width of a character.

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

---

Figure 6. 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.
Experimental Results and Discussion

This section presents the experimental results based on two line segmentation techniques, (a) modified partial projection method and (b) integrating partial projection method with recursion. Palm leaf manuscript images were collected and scanned by Project for Palm Leaf Preservation in Northeastern Thailand Division, Mahasarakham University [9]. The resolution of the input images is 200x200 dpi in RGB format. In this experiment, the set of test images for this comparison includes 43 binary images which are converted from input images by automatically selecting the optimal binarization algorithm in [10].

Example results of this experiment are shown in Figure 7. In the experiment, 195 text lines were segmented from 43 palm leaf manuscripts. To check whether a text line is segmented correctly, a boundary is drawn between two lines. The result of line segmentation is measured by following the rules in [3]. Results of the experiments are given in Table 1.

Table 1. Results from text line segmentation

<table>
<thead>
<tr>
<th>Number of components out from their correct line</th>
<th>Percentage of components segmented within the correct line</th>
<th>Modified partial projection method</th>
<th>Proposed method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Number of correct lines is done</td>
<td>Number of correct lines is done</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Percentage of correct line is done</td>
<td>Percentage of correct line is done</td>
</tr>
<tr>
<td>0</td>
<td>100%</td>
<td>114</td>
<td>137</td>
</tr>
<tr>
<td>1</td>
<td>98.00-99.99%</td>
<td>57</td>
<td>28</td>
</tr>
<tr>
<td>2</td>
<td>96.00-97.99%</td>
<td>10</td>
<td>18</td>
</tr>
<tr>
<td>3</td>
<td>94.00-95.99%</td>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td>≥4</td>
<td>&lt;=93.99%</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

The above technique can be used to separate some touching characters from consecutive lines. However, errors may be due to prolonged characters in each column and being adjacent between two consecutive columns. This has affected a few of the characters in the document. Furthermore, some binary images are unclear and they have a major effect on the accuracy of text line segmentation.

Conclusion and Future Work

This study has compared two line segmentation approaches which are the modified partial projection profile, and the proposed integrating partial smooth and original projection profile with recursion. The proposed technique provided better performance 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, and (2) prolonged components are cut by linking columns. As there are limited reports on the research of line segmentation for Thai and ancient Thai manuscripts, this study will aid 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. Future development will aim at enhancing the performance of the proposed technique by considering the identification of prolonged components. More data sets will be used to test the prototype in order to verify the fully automated knowledge and information extraction system for ancient Thai manuscripts. Furthermore, character segmentation has to be developed to extract the component for recognition systems.

Acknowledgments

The authors wish to express their thanks and appreciation to Mahasarakham University for conference funding, and members of the Preservation of Palm Leaf Manuscripts Project, Mahasarakham University, Thailand for their support and providing images from their database.
Figure 7. Sample results of line segmentation (a) original images, (b) modified projection profile method, and (c) proposed method


6 References


