Image Segmentation of Historical Handwriting from Palm Leaf Manuscripts

Olarik Surinta and Rapeeporn Chamchong

Department of Management Information Systems and Computer Science
Faculty of Informatics, Mahasarakham University
Mahasarakham, Thailand
e-mail: olarik.s@msu.ac.th, rapeeporn.c@msu.ac.th
Telephone: +6643754322 ext 2497
Fax: +6643754359

Abstract: Palm leaf manuscripts were one of the earliest forms of written media and were used in Southeast Asia to store early written knowledge about subjects such as medicine, Buddhist doctrine and astrology. Therefore, historical handwritten palm leaf manuscripts are important for people who like to learn about historical documents, because we can learn more experience from them. This paper presents an image segmentation of historical handwriting from palm leaf manuscripts. The process is composed of three steps: 1) background elimination to separate text and background by Otsu's algorithm 2) line segmentation and 3) character segmentation by histogram of image. The end result is the character's image. The results from this research may be applied to optical character recognition (OCR) in the future.

Keywords: Palm Leaf Manuscript, Image Processing, Image Segmentation, Background Elimination, Otsu's Algorithm

1. Introduction

Palm leaf manuscripts have been a popular written media for over a thousand years in Southeast Asia. [1-3] Palm leaves were used for recording the history, knowledge and local wisdoms such as medical treatments, Buddhist doctrine, astrology and the story of dynasties. There are various texts written on palm leaf manuscripts. [4] An example page of palm leaf manuscript is shown in Fig. 1. . With the passage of time, most of these palm leaves are nearing the end of their natural lifetime or are facing destruction from elements such as dampness, fungus, ants and cockroaches. For this reason, Mahasarakham University is establishing Palm Leaf Manuscript Preservation Project for the discovery, preservation and protection of palm leaf manuscripts from Northeast Thailand. [5]

2. Proposed framework

To extract data from historical handwritten palm leaf manuscripts, the BILAN (palm leaf manuscripts) system is proposed. The user can understand the system by utilizing an easy to use graphical user interface. [5] The system will display image results step by step. Fig. 1 represents the modules within the proposed system.

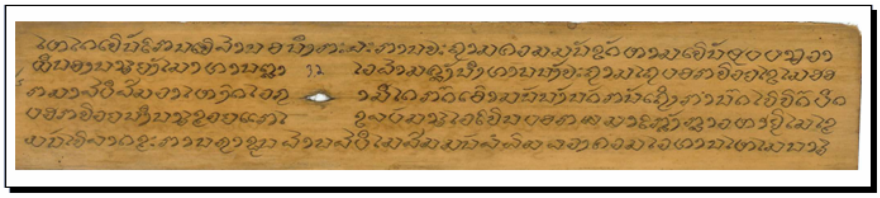


Fig. 1. An example page from palm leaf manuscript.

In our work, we use palm leaf manuscripts consisting of 227 pages to do research work. We implement the system to extract data from palm leaf manuscripts. The system processes consist of background elimination, line segmentation, and character segmentation.

2.1 Convert Image from RGB Color to Grey Image

A RGB color is another format for color images. It represents an image with three matrices of sizes matching the image format. Each matrix corresponds to one of the colors red, green and blue. [1] When we convert it into a grey scale (or "intensity") image it depends on the sensitivity response curve of detector to light as a function of wavelength. [6, 7] The equation is:

$$Y = 0.3R + 0.59G + 0.11B \tag{1}$$

A result from this equation is shown in Fig. 3.

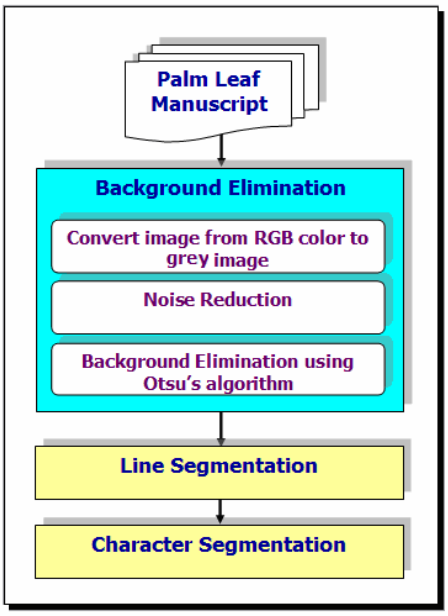


Fig. 2. Framework of the proposed system.

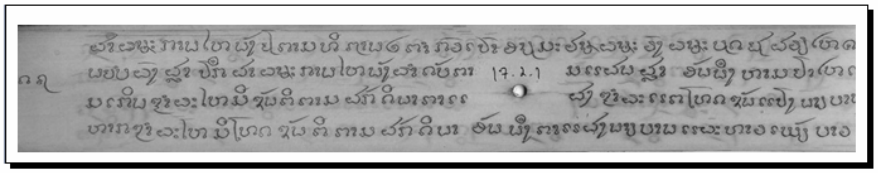
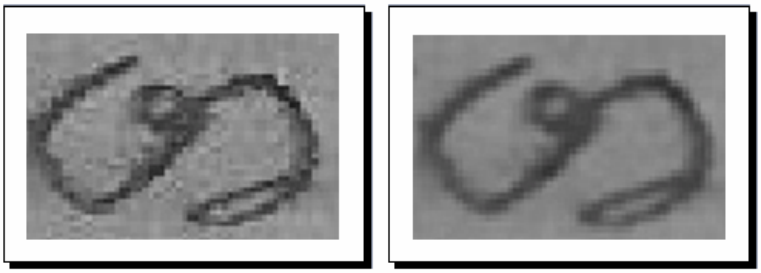


Fig. 3. Figure 1. An example showing a grey image.

2.2 Noise Reduction

Noise reduction is the process of removing noise (from the scanning process) from a signal. A popular technique for removing noise from a grey image is Gaussian filtering. This techniques for calculate the transformation to apply to each pixel in the image. The equation is: [5]

$$f_f(i,j) = \frac{1}{s_k} \sum_{m=1}^{k} \sum_{n=1}^{k} B_{mn} * f_{gr}(m,n) \tag{2}$$



(a) Noise grey image (b) Grey image after reduction

Fig. 4. An example showing a grey image before and after noise reduction.

2.3 Background Elimination using Otsu's Algorithm

The background elimination method proposed by Otsu [5, 8, 9] has the advantage of not needing any prior knowledge of the image, based only on its grey level histogram. The main idea is to find in the histogram an optimal threshold that divides the image objects by constructing two classes from any arbitrary grey level, using the discriminated analysis shown in Fig. 5.

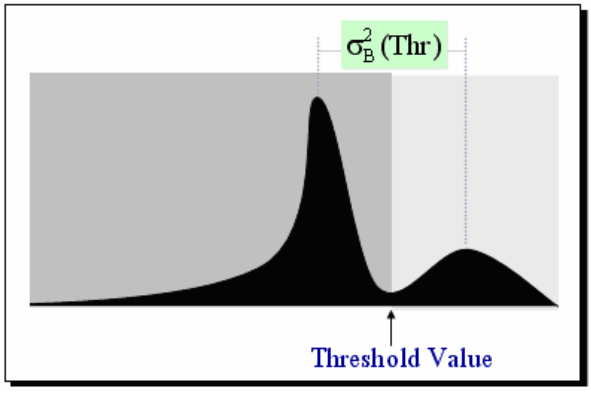


Fig. 5. Otsu's threshold value method.

To find the optimal threshold (*Thr*) we can use the following criteria equation which respects *Thr*.

$$\eta = \frac{\sigma_B^2}{\sigma_{Thr}^2} \tag{3}$$

Where σ^2_{Thr}, that is the total variance, is independent from the grey level, only being necessary to minimize the function σ^2_B, that is the within-class variance. The optimal threshold *Thr** will be defined in the following equation.

$$Thr^* = ArgMin\ \eta \tag{4}$$

A result from this equation is shown in Fig. 6.

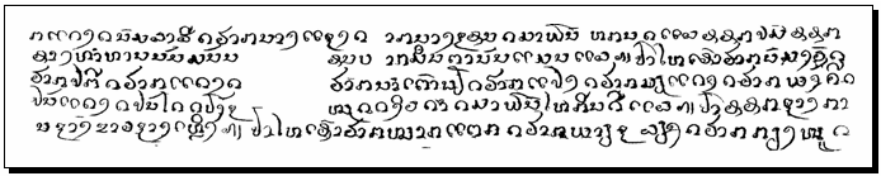
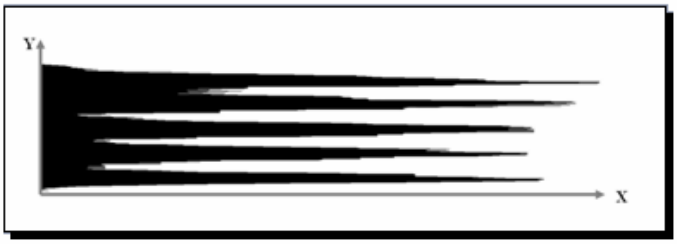


Fig. 6. An example showing a binary image after background elimination.

2.4 Line Segmentation

For the next step, projection profile analysis is a popular technique for line segmentation. We use horizontal projection profile analysis because the texts in most document images are aligned along horizontal lines. The technique computes horizontal projection histograms, the count of black pixels for each column of the raster image. [5, 10, 11]

When the horizontal projection profile is applied on an M × N image, a column vector of size M × 1 is obtained. Elements of this column vector are the sum of pixel values in each row of the document image. An example of the projection profiles of an image is shown in Fig. 7. The peaks in Fig. 7(a), which correspond to the horizontal projection profile of the image.



(a) Line segmentation histogram

(b) Image after line segmentation

Fig. 7. An example showing an image after horizontal projection profile.

2.5 Character Segmentation

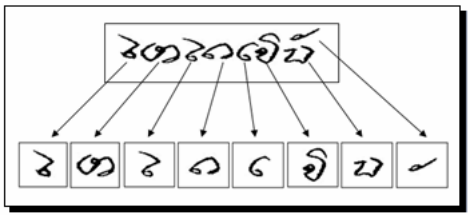


Fig. 8. An example showing an image after vertical projection profile.

As a final step, the extracted lines are segmented into characters. [11-13] To find the boundaries between the characters, we apply a threshold value on the length of the space in between the characters. After finding the positions of the spaces between characters we also eliminate the parts of the line segment. An example character segment is shown in Fig 8.

3. Experimental Results

The method was tested using a set of 227 palm leaf manuscripts. An example palm leaf manuscript is shown in Tables 1 and 2 give an indication of the accuracy.

Table 1. Background Elimination Results

Background Elimination	Accuracy	
	Number of documents	Percentage of background elimination segmented
Complete	138	61
Incomplete	89	39
Total	227	100

Table 2. Line Segmentation Results

Number of Segmented Lines	Percentage of lines correctly segmented
4	78%
5	87%
Average	82.5%

4. Conclusion

In this paper we presented image enhancement techniques for historical palm leaf manuscript document images. Our algorithm first converts the color image into a grey image, then converts the grey image into a binary image using Otsu's algorithm, and finally produces the segmented lines and characters using projection profile analysis.

5. References

[1] Shi Z, Setlur S, Govindaraju V. Digital Enhancement of Palm Leaf Manuscript Images using Normalization Techniques. 5th International Conference On Knowledge Based Computer Systems; 2004 December 19-22, 2004 Hyderabad, India; 2004.

[2] Shi Z, Govindaraju V. Historical Document Image Segmentation Using Background Light Intensity Normalization. 12th SPIE Document Recognition and Retrieval; 2005 January 16-20, 2005; California, USA; 2005.

[3] Shi Z, Govindaraju V. Historical Document Image Enhancement Using Background Light Intensity Normalization. 17th International Conference on Pattern Recognition; 2004 23-26 August 2004; Cambridge, United Kingdom; 2004.

[4] S.A. Shahab, Wasfi G. Al-Khatib, Sabri A. Mahmoud. Computer Aided Indexing of Historical Manuscripts The Inte6rnational Conference on Computer Graphics, Imaging and Visualisation (CGIV'06); 2006 April 7, 2006; Sydney, Australia; 2006.

[5] Chamchong R, Surinta O. Text Line Segmentation from Palm Leaf Manuscripts. The 3rd National Conference on Computing and Information Technology (NCCIT2007). Bangkok, Thailand 2007.

[6] Surinta O, Jareanpon C. Comparison of image analysis for Thai handwritten character recognition. 4th International Conference on Intelligent Information Processing (IIP2006); 2006 September 20-23, 2006; Adelaide, Australia: Springer; 2006.
[7] Surinta O, Nitsuwat S. Handwritten Thai Charactr Recognition Using Fourier Descriptors and Robust C-Protytype. Information Technology Journal. 2006 January - June 2006;2(3):96.
[8] Otsu N. A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man and Cybernetics. 1979 January 1979;9(1).
[9] Otsu's method. 2008 February 20, 2008 [cited 2008 February 20]; Available from: http://en.wikipedia.org/wiki/Otsu's_method
[10] Tacnet LL-SAZB. Text Line Segmentation of Historical Documents: a Survey. International Journal on Document Analysis and Recognition, Analysis of Historical Documents. 2006.
[11] C. Weliwitage, A. L. Harvey, A. B. Jennings. Handwritten Document Offline Text Line Segmentation. In: Cairns Q, Australia 4870, editor. The Digital Imaging Computing: Techniques and Applications (DICTA2005); 2005; Queensland, Australia; 2005.
[12] Ataer E, Duygulu P. Retrieval of Ottoman Documents. 8th ACM SIGMM International Workshop on Multimedia Information Retrieval; 2006 October 26-27, 2006; California, USA: MIR 2006. 8th ACM SIGMM International Workshop on Multimedia Information Retrieval; 2006.
[13] A. Cheung, M. Bennamoun, N. W. Bergmann. An Arabic optical character recognition system using recognition-based segmentation Pattern Recognition Society. 2001 February;3(2).