International Journal on Advanced Science Engineering Information Technology

Vol.9 (2019) No. 6
ISSN: 2088-5334

Research and Development of Feature Extraction from Myanmar Palm Leaf Manuscripts for the Myanmar Character Recognition System

Nwe Nwe Soe[#], Win Htay[*]

[#]*Faculty of Computer Science Department, University of Computer Studies (Thaton), Thaton, 12043, Myanmar*
E-mail: nwenwesoe.mlm@gmail.com

[*]*Principal, University of Computer Studies (Thaton), Thaton, 12043, Myanmar*
E-mail: wynnhtay@gmail.com

***Abstract*— This paper proposed Myanmar palm leaf manuscript handwriting OCR system. Each text area in the Myanmar palm-leaf manuscript is segmented. This segmented character text image is needed to be recognized to transform to Myanmar handwritten characters which express Myanmar's precious historical and invaluable information. This paper involves two essential steps: preprocessing and feature extraction. The preprocessing is carried out to extract the attractive palm-leaf manuscript region from the Images automatically are taken by the camera and to support the enhanced images for subsequence processes of Myanmar character recognition from Myanmar palm leaves. The one-dimensional segmentation approach is used to crop leaf area in the image which is taken with high resolution. Line count analysis is also done to extract the region for using enough line count. After that, line segmentation is carried out using Object Frequency Histogram along the horizontal lines which can find the best optimal points between the lines. Similarly, the same technique but vertically is used to get each character or smallest group of characters. Totally 18 features are extracted to recognize the Myanmar palm-leaf manuscript characters. Although the experimental results are good enough but some difficulties are still needed to take account related to the connected components.**

***Keywords*— Myanmar palm leaf manuscripts; feature extraction; OCR; line segmentation; character segmentation.**

I. INTRODUCTION

U Phoe Thee Buddhist Scriptures Library situated in the compound of Sa-dama-zawtikar-yarns Monestory is at the foot of Mya Tha Bake hill in Thaton, Mon State, Myanmar. A retired mayor established it U Boe Thee, who donated his treasured collections of Buddhist manuscripts and Parabaiks for the monks to be able to study. It has now been undertaken by an organization called Suwanabumi Pari Yutti Sarsana Philadelphia Association. The ancient palm-leaf inscriptions, Parabaik and other Pitakas written in such languages as Mon, Myanmar & Pali are saved in a big cupboard enshrined with real gold. Moreover, one can also study many other scriptures dating back from 130 to 250 years ago. Therefore, this Library is of invaluable treasure resource for international researchers as well as for those who have a strong interest in Pali, Mon and Myanmar languages.

In Myanmar, there are many organizations and institutions committed for the protection of ancient palm leaf manuscripts in order to store precious knowledge writings. One such objective is to save the precious palm leaf manuscripts images for effective retrieval of valuable data automatically from these manuscripts. Many image processing techniques have been proposed for the efficient data retrieval from manuscripts [1].

The palm leaf manuscripts are the important sources of the historical events of the world. In the area of traditional Medical Affairs, Mathematics, Literature, Myanmar culture life style, many researchers refer the manuscripts for their research work. In Myanmar, one can find digitized valuable palm leaf manuscripts nowhere except in Pali text Society at U Phoe Thee Library where a need job of digitizing palm leaves can be seen [2]. It is time now to preserve and save valuable treasures of our nation and it will be beneficial if they are saved in digital format files or on the Web sites.

In addition, texts from these palm leaf manuscripts need to be extracted or rewritten in the form of digital character. The aim of the proposed research is to help the work of digitizing these valuable palm leaf manuscripts stored in the different places, such as Myanmar National Library, Yangon University Library etc. U Pho Thee Library is a famous Myanmar ancient manuscripts library is situated in Thaton, Mon State, Myanmar. This library has packed over 788 Parabaik manuscripts and ancient books. These scrolls

religious books are always taken care of by Tharduzana Parthardika Thuwana Bume Buddhist Society. It is a must for Myanmar citizens to maintain these Myanmar valuable palm-leaf manuscripts. The fungi on the surface of palm leaf manuscripts can be cleaned by mixing lemongrass oil with carbon. This step is called verification and is necessary to clear the unwanted things over palm leaf manuscripts. After verification, there is a need to take photos of purified palm leaf manuscripts using the high-resolution digital camera [2].

The objective of the Myanmar Palm-leaf manuscript system is to improve the quality of the documents before a recognition system uses them for subsequent information feature extraction. The palm leaf manuscripts are the most important sources of the historical events of the world. In the area of traditional Medical Affairs, Mathematics, Literature, Myanmar culture life Buddhist laws and so on, many researchers refer to the manuscripts for their research work. In the religious field of Myanmar country, these Buddhist laws are typed and edited easily for reference books. Especially, Myanmar antique Medical, Commentaries on Buddhist Pali texts, Poem of epic proportions, Prosody, Kind of four standard verse, Compilation of learned discourses, Code of Laws, Myanmar Mathematics and Cabalistic treatises of Palm leaf manuscripts are so complicated as well as precious and thus they must be maintained and converted to digital format files.

Up to now, no research work about Myanmar palm-leaf manuscript OCR is done. Before this work, previous research as the preprocessing is already carried out to support the enhanced images for subsequence processes of Myanmar character recognition from Myanmar palm leaves. A high-resolution digital camera takes these Myanmar palm leaf manuscripts images with Jpeg format file system not a scanner. In [1], an advice can be set to define the background color in the Image Acquisition process for Myanmar palm leaves by analyzing research output. Moreover, an analysis [3] is also done to define only the Red color channel for the binarization step in the Myanmar palm leaf processing.

II. Materials and Methods

A. Materials

The past OCR system for various scripts of the literature numbers and several researches related to Myanmar Palm leaf manuscripts OCR system are reviewed. Nowadays, many researchers tested the printed and other handwritten OCR in Myanmar, but no researcher has ever tested Palm leaf handwritten manuscripts OCR system. Data collecting and photographing processes for these Palm leaf manuscripts have not been completed yet.

So, the authors are trying for their experiment databases comprising samples of characters from the Myanmar palm leaf document image. There is a collection of over 200,000 Myanmar handwritten palm leaf document images readily for the image processing acquisition research work (Myanmar, Mon and Pali languages) from the manuscript library of U Phoe Thee.

B. Methods

1) Myanmar Characters and Palm leaf manuscripts: In Myanmar, Pali and Mon characters, there is no difference between Upper Case and Lower-Case characters. The direction of writing style is from left to right, moving hands horizontally. Palm leaf manuscripts had typically been in use last few centuries but over time, the palm leaves gradually degrade in quality and they are neither popular nor useful in any form. In this research, extracting text for digital images of Myanmar palm leaf manuscripts OCR system is presented. The resolution of the input images in the Jpeg format photos is illustrated in the following Figure 1.

Fig 1: Difference RGB color intensities values between Leaf area and Various Background area (Normal, Black, Green and Blue)

2) Nature of Myanmar Manuscript: Myanmar scripts were written in a variety of languages like Myanmar, Mon, Pali etc. The script is written from the left side to the right side and there is a header line that combines each character to form words. There are upper and lower modifiers that make the segmentation task more difficult. There are 35 consonants, 5 combining marks, 6 symbols and punctuations, 7 independent vowels, and 10 digits and 8 vowels present in Myanmar characters. There may be two or more consonants that can be composed together to upper, lower or right characters. Basic Myanmar scripts are shown in Figure (2) and (3).

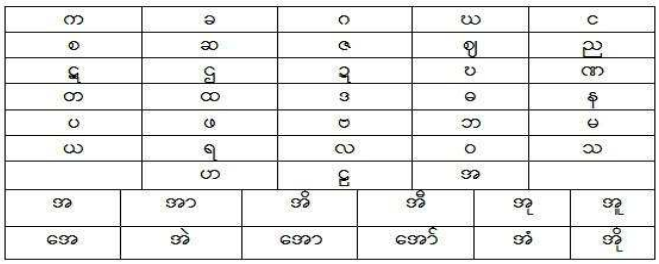


Fig 2: Basic Myanmar characters and vowels

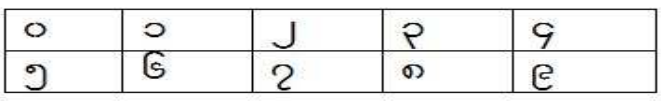


Fig 3: Myanmar digits

3) Steps of feature extraction

- Cropping leaf area from the images (get cropped image) using 1 Dimensional Segmentation.

This method is based on the intensity value function along the line and it is more suitable because of its more accurate values. If only one line uses the least accuracy and it is necessary to test with the suitable (fair) lines to find the edges with differential value of Intensity function along the line that is the best results to crop the Palm leaf manuscripts areas as shown in Figures 4(a), (b) and (c) respectively.

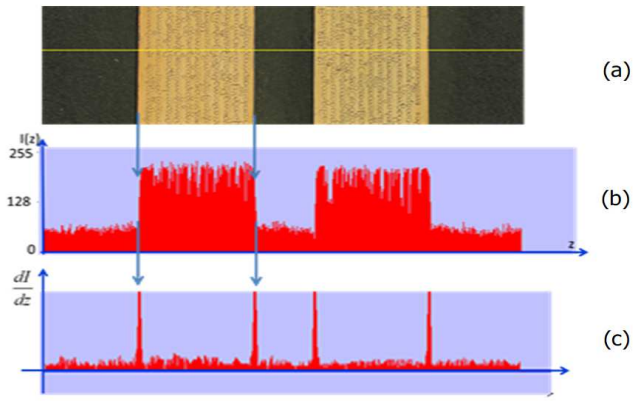


Fig 4: (a) A segment of Palm leaf image with a line (b) Intensity function along the line (c) differential value of Intensity function along the line

The one-dimensional segmentation approach is that object location can be determined by clustering points of interest and hierarchically forming candidate of palm leaf manuscript regions according to similarity and spatial proximity predicates, as shown in Figure 5. This system can be used to optimize two factors: RGB background colors and the number of vertical lines to choose candidate areas. Moreover, one-dimensional edge segmentation performs better accuracy and less calculation time than other traditional filters. This system can be applied to the segmentation of the candidate area which includes text. The results of the research can be used as an input image to implement an OCR system to provide information on being existence for their related fields [4].

-2/7	-1/7	0	1/7	2/7

Fig 5: Improved one dimensional edge segmentation mask

Traditional edge detection methods such as Sobel, Canny and other related edge detectors needed to use more calculation steps and time. As shown in Figure 6, edge detection is carried out only across the line, top to bottom of the image. Too successfully crop the leaf, five lines are adequate for edge detection. So, the proposed one-dimensional edge detection method is very simple and takes less time than other edge segmentation methods.

Line count analysis is done to extract the region of interest from the various background color types. According to the result below, the blue background is the most suitable and we found that the number of lines counts greater than five is the best and exactly accuracies. So later in Myanmar photographing and collecting of palm leaf manuscripts processes, an empirical method and a novel contribution is proposed, which is taking a photograph with a blue background color which is automatically cropped the image with more accurate results.

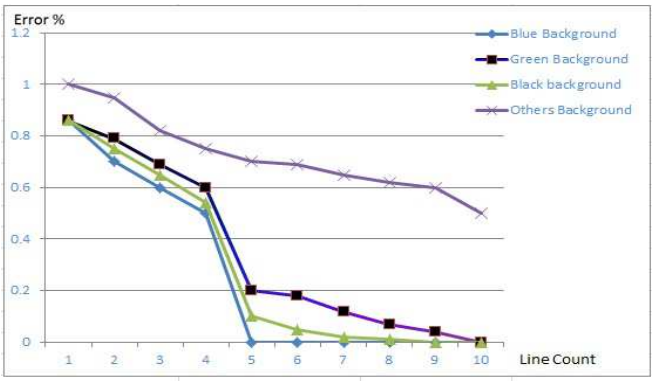


Fig 6: The Error % line counts for various backgrounds graph

- Line Segmentation (Projection, histogram, Otsu).

In [5], an approach for the segmentation of characters from ancient Malayalam palm leaf manuscripts were presented using cognitive memory networks. Survey on Text Line Segmentation of Historical Documents is presented in [6]. After cropping, the next step is the binarization process. In this paper, an empirical method is carried out to know what color intensity is the best to use for binarization. Four binary output images are shown in the following figure. Figure 7 (a) shows binary image getting from Red color channel, and other Figures 7(b), (c) and (d) show the binarization results using Green, Blue and Gray color channels respectively.

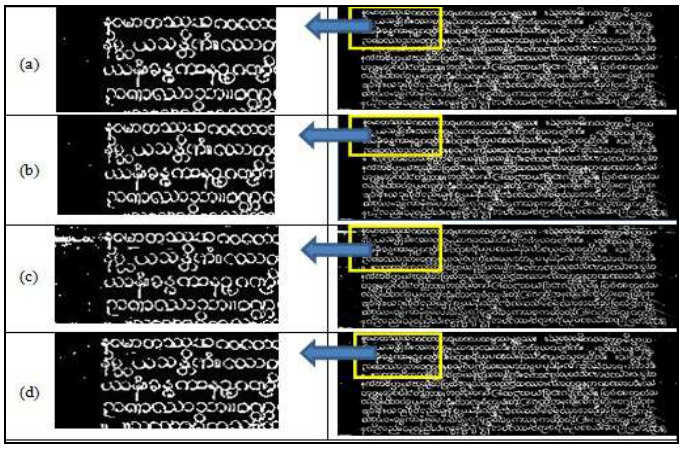
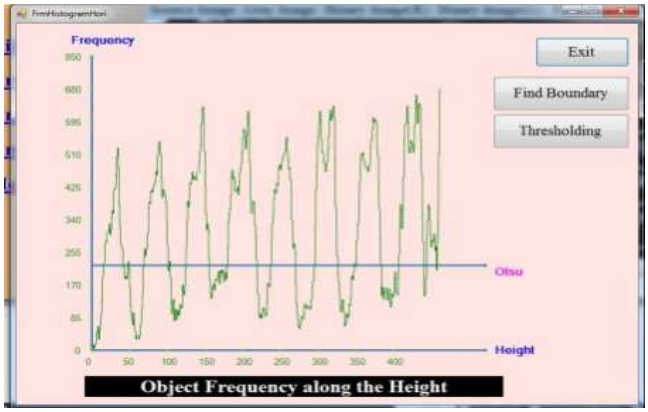


Fig 7: Binary Images: (a) From Red Color Intensity, (b) From Green Color Intensity, (c) From Blue Color Intensity, (d) From Gray Color Intensity

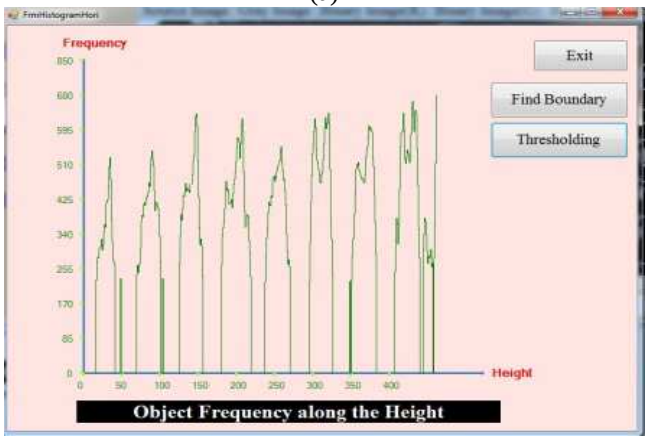
Binarization result processed using Red color channel is the best for the binarization process. There is no need to consider other channels and change the grey image for the binarization step. For the palm leaf, this empirical method is more suitable as the processing time is less than the other traditional methods. The output binary image is used for the next step, which is line segmentation. Firstly, the optimal points between the lines are searched. Lines are segmented along with these optimal points. To extract each character image or smallest group of a character image, firstly in this paper, image segmentation line by line is done using object

frequency along the horizontal lines. Object frequency histogram is used to find the best optimal points.

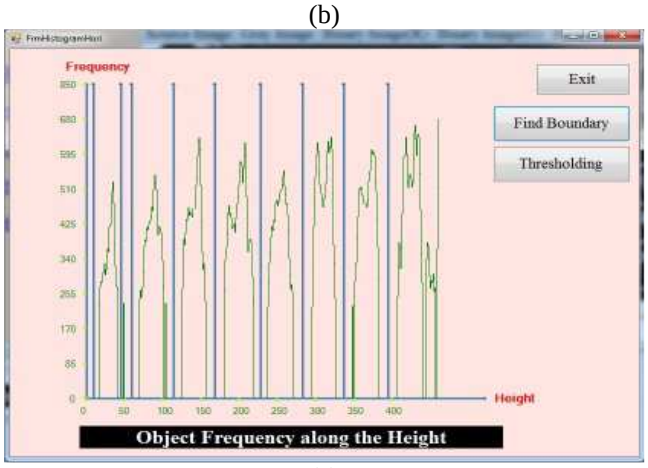
In Figure 8, the object (character) pixels frequency histogram is shown. In Figure 8(a) shows how to find the white pixel object frequency along the width of the image and the blue line is the threshold value by using the Otsu thresholding algorithm to remove the weakest frequency. Figure 8 (b) illustrates the histogram after thresholding and the possible candidate locations of the line segmentation positions. In the last Figure 8 (c), borders of the lines which are the middle points of possible spaces between two-character lines are defined [3].



(a)



(b)



(c)

Fig 8: (a)Histogram of object frequency along the width of the image, (b)Histogram after thresholding, (c)Searching the optimal points between the line areas

The line segmentation process is shown in the following Figure 9.

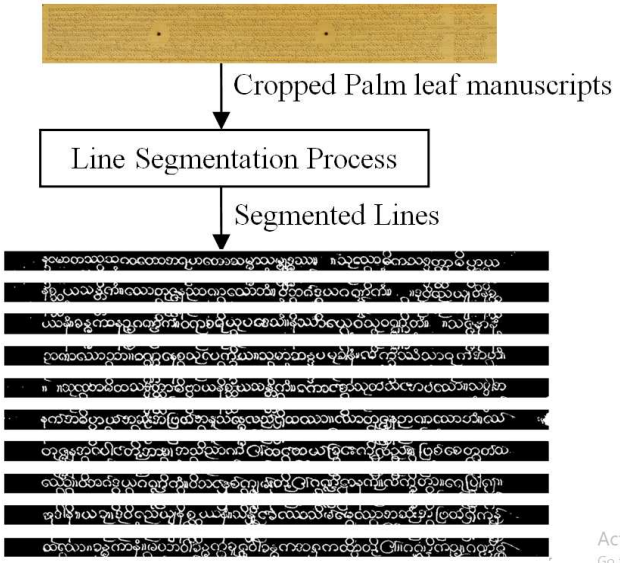


Fig 9: Line segmentation from palm leaf manuscript

- Character Segmentation.

Similarly, to carry out character segmentation, the same technique is used as a line segmentation process, but vertically. After the line segmentation process, each line containing characters is the input to the next step, character segmentation. The input image is convolved vertically to find the borderline of each character. There are spaces between the syllabus and each character. Nevertheless, in the handwritten documents, consideration only about spaces for character segmentation is not enough. There are noises and connections between characters. The character segmentation process is shown in Figure 10.

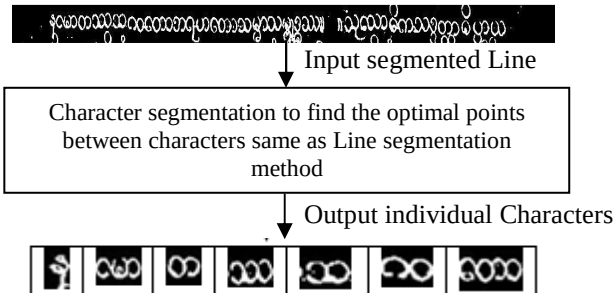


Fig 10: Character segmentation process from line

- Feature Extraction of Myanmar Characters

The 3D feature-based character recognition system for palm leaf manuscripts are presented in [7] with an accuracy of 96%. Soumya A. and G. Hemantha Kumar [8] presented work on Fourier feature-based classifiers for recognition of Kannada epigraphs. This method performs feature extraction in the first step, then global recognition was performed by comparing the representation of the unknown word with the references stored in the lexicon. Consequently, this method uses the “classical approach”, with complete words as symbols to be recognized. For instance, the scale-space technique [9] and holistic word recognition [10] are normally used on Roman scripts. These methods are

restricted in application to a predefined lexicon. Feature extraction is an essential step for Myanmar Character Recognition System. Myanmar character features are not like English and other languages. In this work, these features are considered for the Myanmar OCR.

In this work, 18 features are extracted for Myanmar OCR system and they are as follows:

- Feature 01: The ratio of Object pixel and Total pixel.
- Feature 02: The ratio of Width and Height of segmented character object image. (Shape)
- Feature 03: Global CG
- Feature 04: Left openness percentage
- Feature 05: Right openness percentage
- Feature 06: Upward openness percentage
- Feature 07: Downward openness percentage as shown in Figure (11).

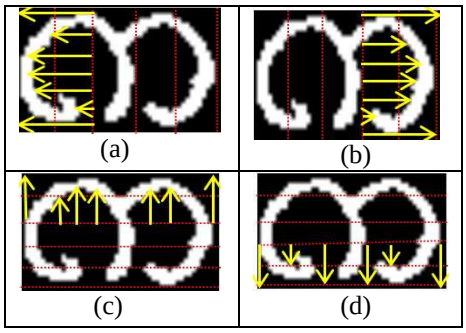


Fig (11) :(a) Left, (b) Right, (c) Upward, (d) Downward Open % of Myanmar Character image

- Feature 08: Number of holes. Myanmar character images include holes which can define the Myanmar characters. So, the count of holes or loop is one of the key features for Myanmar OCR. Example of separated characters having no hole, one hole, two holes, three holes, and four holes is shown in the following Figure 12.

Fig 12: Zero hole, One-hole, Two holes, Three holes and Four holes

- Feature 09: Global Orientation. The orientation of the average direction of all the pixels in the image.
- Feature 10-13: Four Local Orientations. Image is divided into many parts and finding the orientation of each part to get more actuate results are illustrated in Figure 13. In this paper, only four local orientation features are calculated: Upper left, Bottom left, Upper right and Bottom right.

Fig 13: Local Orientations

- Feature 14-17: Local CG (4 paths). The Local CG divides the value to four quadrants of Top left, Top right, Bottom left and Bottom right and then it is necessary to find the desired CG of these four partitions.
- Feature 18: Object location. This feature is calculated not from the segmented image but from the source image, which is the input for the character segmentation process. In Myanmar characters, this feature is important to distinguish the Myanmar character. This feature is considered in which level it is located: Upper, Middle and Lower levels are illustrated in Figure 14.

Fig 14: Three levels of Myanmar characters

The most prominent object in this figure is in the middle and lower level. The location of Global CG is used to define in which level it falls. The labeling method is used to extract the character area from the image. If the area of this detected character is extracted using maximum X, Y and minimum X, Y, other object areas will be included to the extracted area. In this case, all other labels are considered as the background.

It should be assumed that foreground object pixels are the white and background are black. To find the white pixel (Object) it is necessary to go from top to down and left to right. If an object pixel is found, all the connected pixels to this are given a label. So "**Kagyi" character** is firstly detected as in the following Figure 15(b) and (c). And then, Thaethae Tin character is detected in Figure 15 (d) and (e).

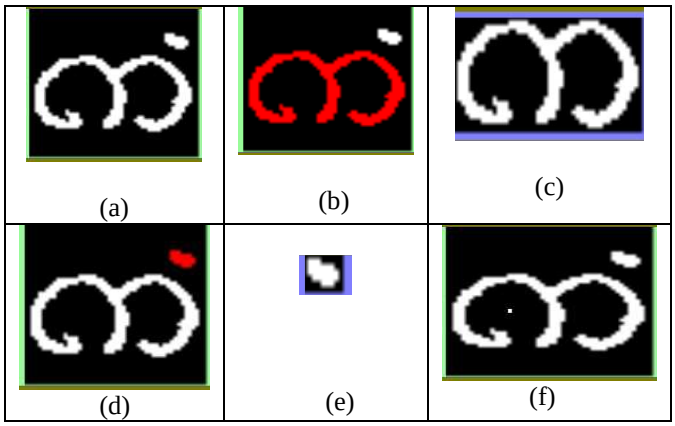


Fig (15): (a) Kan Myanmar word original image, (b) The red color cropped Kagyi image, (c) The red color cropped Thaethae Tin character image, (d) Lower than 10 white connected pixels removed from image as a noise

Very Small-Object which has less than 10 pixels been defined as the noise and removed. An example of this type of object is shown in Figure 15 (f). An application is developed to carry out the feature extraction process. After extraction of each character, there is a need to name this character and save all the features of it with the given name for the next step of Myanmar OCR. The user interface of the developed system is illustrated in Figure 16.

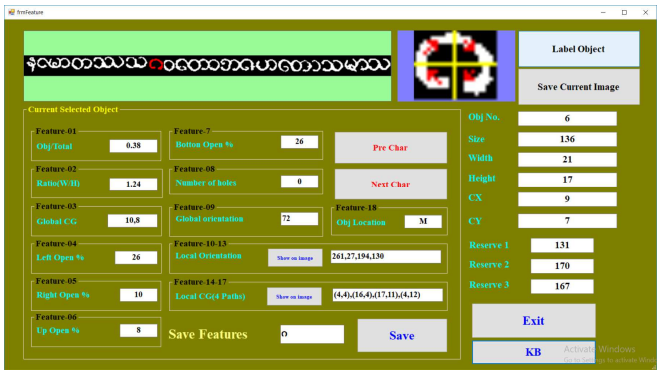


Fig 16: Feature extraction outputs

Extracted features (Feature 01 to Feature 18) are good enough and suitable for the Myanmar OCR recognition system. Up to now, there is no researcher working for the Myanmar palm leaf manuscript OCR system in our country. As the discussion with the experts, these features are sufficient and the best for the Myanmar character features from the experts' point of view.

III. Results and Discussion

In fact, this paper is the part of a research work which is text extraction from the Myanmar palm leaf manuscripts. This research is also intended to use in the real world, in the Oo Pho Thee Library, where there are about three hundred thousand digitized leaves: JPEG files nearly 4000 x 500 in size. In this work, the extraction of leaf area from the taken photos was introduced. And for the time being, the extraction leaf area has been already done for all the leaves using the illustrated method.

Line and character segmentation processes are also successfully carried out. The results showed that character segmentation has correctly been done. Feature extraction work has now been carried out using these segmented characters and features of each segmented character are saved in the database to use in the next step of the Myanmar OCR system. In this paper, 18 features are calculated for the recognition process. Experts in the area of OCR think that these features are just enough for the Myanmar OCR system. In the developing process, there had been so many difficulties such as writing style, smearing from ink, the direction of the line, touching and noises. Deviation of Line direction, slant position or line skew is the big problem for the line segmentation as seen in the following figures 17 and 18.

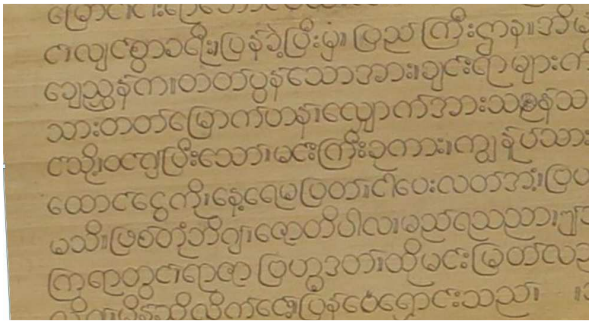


Fig 17: Example slant image

The proposed technique is derived from the partial projection method [11]-[16]. The text image is divided into vertical columns and then the histogram of the projection profile is applied by smoothing [14] to separate the lines. Smoothing is used to remove spurious peaks and valleys of the histogram.

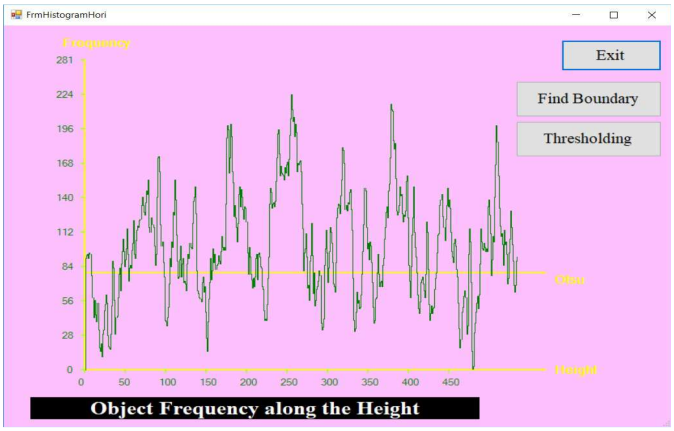


Fig 18: Output Histogram of object frequency along the width of the image showing difficulty to find the optimal points between the line areas

The Up, Down, Left and Right touching components may be the next big problem is the result of character segmentation. The touching components are difficult to separate due to the writing style and smearing from ink just because of using the labeling method to extract characters. Example output of the character segmentation process showing connected characters is illustrated in the following Figure 19.

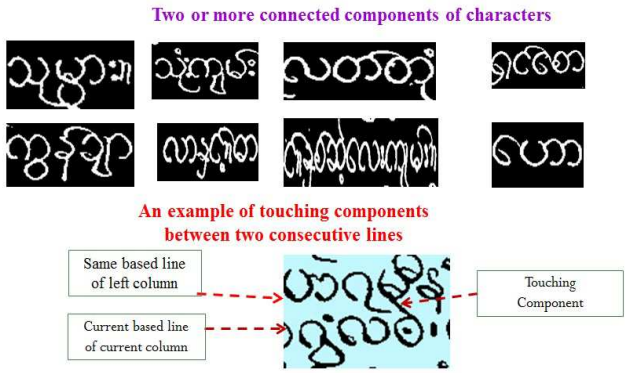


Fig 19: Touching problem between two lines

If the segmented character is longer than the reasonable length of Myanmar character, this character needs to be divided again. So adaptive segmentation must be done for these characters. More features can be added for better accuracy if it is required to divide more blocks to find Local CG and Local Orientation features.

IV. Conclusion

This is the first Myanmar Character Feature extraction from the Myanmar Palm Leaf. From the experts' point of view, these features are sufficient and enough for the Myanmar OCR palm leaf handwritten character recognition system. The empirical methods are used to extract the leaf area and to decide which color channel should be used for

the binarization process for time-saving. Experimental results show that the proposed system is a reliable and robust technique for character segmentation and feature extraction process. Features are now saved in a database and the authors are still carrying out further work of character recognition for these stored features.

ACKNOWLEDGMENTS

The authors would like to express their heartfelt thanks to the Palm Leaf Manuscripts Project, Oo Phoe Thee library, Thaton, Myanmar, for their collecting and supporting images from the database.

REFERENCES

[1] Alahakoon, C. N. K., “Identification of physical problems of major palm leaf manuscripts collections”, Sri Lanka. J. Univ. Libr. Assoc. Sri Lanka, 2006, October, pp.54–65.

[2] Nwe Nwe Soe, Win Htay, “Finding region of interest and automatic cropping from Palm leaf manuscripts by using one-dimensional segmentation”, 14th ICCA Conference, 2016, February.

[3] Nwe Nwe Soe, Win Htay, “Syllabus segmentation from Palm leaf manuscripts”, 16th ICCA Conference, 2018, February.

[4] Nwe Nwe Soe, “Syllabus Line Segmentation from Palm Leaf Manuscripts by using Vector Neural Network”, Journal of Applied Informatics and Technology (JIT), Thailand, 2018, Volume-1, Number 1, January – June.

[5] Kumar, Neethu S., Dwivedi Sanjeet Kumar, S. Swathikiran, and Alex Pappachen James. "Ancient Indian document analysis using cognitive memory network." In Advances in Computing, Communications and Informatics (ICACCI, 2014 International Conference on, 2014, pp. 2665- 2668. IEEE.

[6] Likforman-Sulem, Laurence, Abderrazak Zahour, Bruno Taconet. "Text line segmentation of historical documents: a survey." International Journal of Document Analysis and Recognition (IJDAR) 9, No. 2-4, 2007, pp.123-138.

[7] Lakshmi, T. R., Panyam Narahari Sastry, Ramakrishnan Krishnan, N. V. Rao, and T. V. Rajinikanth. "Analysis of Telugu Palm Leaf Character Recognition Using 3D Feature." In Computational Intelligence and Networks (CINE), 2015 International Conference on, 2015, pp. 36-41. IEEE.

[8] Soumya, A., G. Hemantha Kumar. "Fourier Features for the Recognition of Ancient Kannada Text." In Computational Intelligence in Data Mining—Volume 1, 2016, pp. 421-428, Springer India.

[9] R. Manmatha and J. L. Rothfeder, "A scale space approach for automatically segmenting words from historical handwritten documents," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 27,2005, pp. 1212-1225.

[10] V. Lavrenko, et al., "Holistic word recognition for handwritten historical documents", in Document Image Analysis for Libraries, 2004. Proceedings, First International Workshop on, pp. 278-287.

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

[12] 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.

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

[14] 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.

[15] 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.

[16] 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.