Finding region of interest and automatic cropping from Palm leaf manuscripts by using one-dimensional segmentation

Nwe Nwe Soe, Win Htay
Computer University, Thaton
nwenwesoe.mlm@gmail.com, wynnhtay@gmail.com

Abstract

This paper offers finding the region of interest and automatic cropping from palm leaf manuscripts developed by using one-dimensional segmentation. This system can be used to provide for enhancing palm leaf manuscripts are taken by high quality resolution digital camera. 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. This system can be used to optimize two factors: RGB background colors and number of vertical lines to choose candidate area. Moreover, one-dimensional edge segmentation performs better accuracy and less calculation time than other traditional filters. This system can be applied to segmentation of 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 of being existence for their related fields.

Keywords: Palm leaf manuscripts, one-dimensional segmentation, Region of interest, automatic cropping

1. Introduction

Palm leaves were one of the earliest forms of writing media and their use in the South and Southeast Asia has been recorded from as early as the fifth century B.C until as recently as the late 10th century. Palm leaf manuscripts relating to Art and Architecture, Mathematics, Astrology, and Medicine dating back several hundreds of years are still available for reference today and thanks for many ongoing efforts for preservation of ancient documents by libraries and universities around the world.

Palm leaf manuscripts have attracted rarely research in recent times in Myanmar country and received extensive attention in academic and production fields [1]. It is an important area in image processing and pattern recognition. Document Image Processing is one of the key application areas of image processing. Several research works have been focusing toward evolving newer techniques and methods that would reduce the processing time while providing higher recognition accuracy.

The main cause of stains on the surface of old manuscripts is fungal growth and wound area. Most of the stains, smears and spots are bluish, greenish, blackish and brownish in color, dust particles and insect remnants also cause stains on the palm leaf manuscripts. Improper and inadequate manual handling also adds to stains. Sometimes the thread used

for tying the folios together and cloth used to protect palm leaves also leave stray marks that contribute to stains [2].

U Pho Thee Library is a famous Myanmar ancient manuscripts library and situated in Thaton. This library has the leaves over 200, 000 Palm leaf manuscripts, packed of over 788 Parabaik manuscripts and ancient books. These scrolls religious books always take care from Thuwana Bume Buddhist Society [8]. Especially we are Myanmar citizens so we must take care of these Myanmar valuable palm leaf manuscripts. The fungi on the surface of palm leaf manuscripts can be cleaned by mixing of lemon grass oil with carbon. This step is verification and to make and clear the unwanted things over palm leaf manuscripts. After finishing, we can take the purified palm leaf manuscripts with the high resolution digital camera [8].

1.1. Motivation

The palm leaf manuscripts are most important sources of the historical events of the world. In the area of traditional Medical Affair, Mathematics, Literature, Myanmar culture life style etc. and many researchers refer the manuscripts for their research works. In Myanmar, no one and nowhere digitized valuable palm leaf manuscripts but Pali-text Society at U Phoe Thee Library does a need job, digitizing palm leaves. It is time now to preserve and save valuable treasures of our nation and need to save them in digital format files or Web sites. In addition, these palm leaf manuscripts need to extract text from them or rewrite in the form of digital character. The aim of the proposed paper is to help the work of digitizing these valuable palm leaf manuscripts stored in the different places, such as Myanmar Nation Library, Yangon University Library.

The performance of subsequent steps in document image analysis heavily depends on the performance of ROI and automatic data cropping which are the most important steps of the palm leaf manuscript segmentation. An efficient method for increasing the generalization capacity of edge lines detection is more efficient than other traditional edge methods segmentation by using one-dimensional edge lines segmentation mask. The methods are applied to palm manuscripts photos of black, green and blue background colors. As a sample, black background image of palm leaf manuscript is illustrated in figure 1.

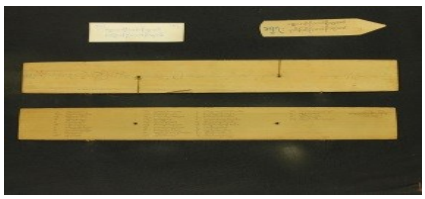

Figure 1: Black background palm leaf manuscript image

2. Related work

There are many algorithms used for image segmentation, and some of them segmented an image based on the object while some can segment automatically. Nicolas and Carole described a method region of interest image segmentation by using fast method but it lacked accuracy [3]. One-dimensional segmentation method is to find region of interest by vertical line counts method to be exact with the weak point of performance of the other systems [4]. In alternative approaching method, current research work is designed to optimize not only accuracy but also performance as necessary.

In [9], a similarity close measure was used to classify the belonging of the pixels, and then used region growing to get the object. The most

common method employed is to select a threshold for the intensity of the image and then convert all the intensity values above the threshold to one intensity value and all intensity values below the threshold value to the other chosen intensity. Unfortunately, it required a set of markers, and if there is an unknown image, it is hard to differentiate which part should be segmented. In [10], a two-step approach to image segmentation is reported.

Edge detection is a very important low-level image processing operation, which is essential in order to carry out various higher level tasks such as motion and feature analysis, understanding, recognition. Over four decades of research work on edge detection have resulted in the development of a plethora of simple to complex techniques. However, as stated in [11–12], the problem of finding the edges has not been solved in its entirety and no universally accepted technique exists.

2.1. Optimization work

Digital photography is a common technology for capturing and archiving images due to the decreased costs for multimedia recording and storage devices, high transmission rates, and improved compression techniques, the digital image collections have grown rapidly in recent years. This paper intends to improve performance and accuracy of segmentation for candidate area of text. To optimize two factors: background colors (Red or Green or blue or others) and number of vertical lines to choose candidate area for palm leaf manuscripts by using one-dimensional segmentation.

3. System Overview

The system overview of the proposed ROI and automatic cropping is performed as shown in figure 2. The user chooses the desired images for finding ROI and automatic cropping. When the palm leaf image is entered, the system makes preprocessing steps such as noise filtering and so on. After removing noises images, the system performs edge detection step. In edge detection step, the text area are detected by combining own created one-dimensional segmentation method and horizontal or vertical edge lines detection method. These unwanted text areas are cropped automatically by using vertical lines edge detections method.

Firstly, the image preprocessing module inputs palm leaf manuscript images from a user. Secondly, the input image is detected into suitable RGB vertical or horizontal edge detection. Thirdly, the system finds (ROI) regions and then automatically splits into new two palm leaf manuscript image files from original palm leaf manuscript image file by using one-dimensional segmentation method. Finally, the cropped images are saved into the Pdf format file creation system.

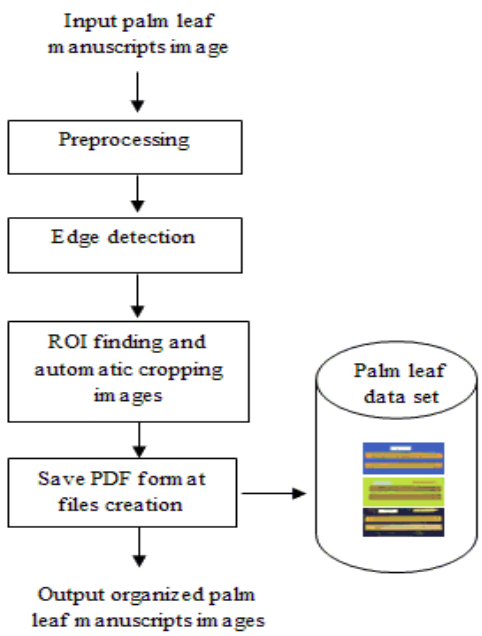

Figure 2: System overview of ROI and automatic image cropping in palm leaf manuscripts system

The system of this research output organized palm leaf manuscripts images is an automatic image cropping system by using enhanced one dimensional edge lines detection mask. This system converts to (*.pdf) file format creation for palm leaf manuscripts library.

4. Region of interest and automatic cropping images

4.1. Pre-processing for palm leaf manuscripts system

This Pre-processing is the name for operations on images at the lowest level of abstraction and it aim is an improvement of the image data that suppress undesired distortions or enhances some image features important for further processing. It does not increase image information content. This method uses the considerable redundancy in images. Neighboring pixels corresponding to one object in real images have the same or similar brightness value. If a distorted pixel is removed from the image, it can be restored as an average value of neighboring pixels [3].

There is difficulty in preprocessing of palm leaf manuscript system. Original condition of the manuscripts is not clean, so that images of digitations result no clear. Therefore, colors which can differentiate objects from backgrounds are sometimes unclear [4]. Manuscript digitations process is imperfect. For example, due to low lighting, manuscript images are not obvious. In addition, manuscript position in digitization is not maximal because the manuscripts can't be opened widely. This system is taken input palm leaf manuscript images with the high quality digital camera resolutions of 4272x2848 pixels.

4.2. One-dimensional segmentation mask

There are three steps to perform edge detection: Noise reduction, Edge enhancement and Edge localization. Noise reduction, where we try to suppress as much noise as possible, without smoothing away the meaningful edges. Edge enhancement, where we apply some kind of filter that responds strongly at edges and weakly elsewhere, so that the edges may be identified as local maximum in the filter's output. Edge localization, where we decide which of the local maxima output by the filters edges and which are caused by noise [5]. The average of the pixels over a region is analogous to integration, differentiation can be expected to have the opposite effect and thus sharpen an image. Edges are pixels where brightness function changes abruptly. Three tristimulus values T1, T2, and T3 can be used to quantify of Red, Green and Blue colors [7].

$$G(j,k)=\left\{\left[G_1(j,k)\right]^2+\left[G_2(j,k)\right]^2+\left[G_3(j,k)\right]^2\right\}^{\frac{1}{2}} \quad (1)$$

Each gradient in the equation (1) represents the three tristimulus component values. A color edge is detected if the gradient exceeds the defined threshold. In the presence of noise every small fluctuation in the intensity value will be amplified by differentiating the image [6]. It is common sense then that we should smooth the image first with low pass filters and then find the local differences.

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

Figure 3: **Improved one dimensional edge segmentation mask**

The proposed system detects the signal by replacing each intensity value with the average of three successive intensity values. Then we estimate the derivative at position *x* to average the two differences between the value at position *x* and its left and right neighbors pixel RGB values. The one-dimensional segmentation mask is illustrated obviously in figure 3, it results as the better smoothing images. The 1st and 2nd derivative of edge detections are illustrated in figure 4.

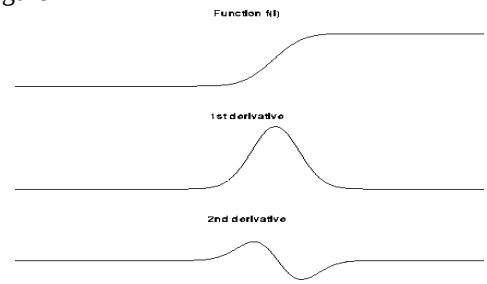

Figure 4: **1st and 2nd derivative of an improved edge line detection**

4.3. Finding ROI of palm leaf manuscripts

The proposed methodology of determining a threshold in a gradient histogram is deduced rigorous analysis and it helps in achieving consistently appreciable edge detection performance in figure 5(a), (b) and (c).

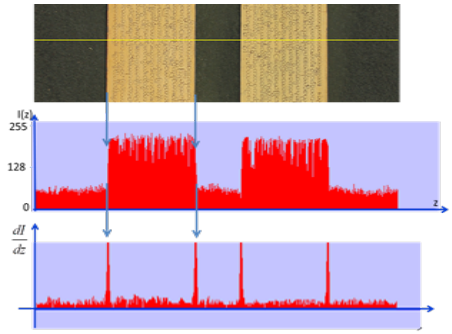

Figure 5: (a) A segment of Palm leaf image with a line (b) Intensity function along the line (c) differential value

5. Experimental results

In this experiment results, palm leaf manuscript images are collected by U Pho Thee library in Thaton city, Mon state of Myanmar country. The proposed system has been implemented with Visual C# and Microsoft Access 2010 Database system. The palm leaf manuscripts image backgrounds are obtained by the Black, Green, Blue and others backgrounds in figure 6.

Figure 6: RGB color intensities values of Normal, Black, Green and Blue background images

Traditional edge detection methods such as Sobel, Canny and other related edge detectors needed to use more calculation steps and time. The proposed one-dimensional edge detection method is very simple and less time than other edge segmentation methods.

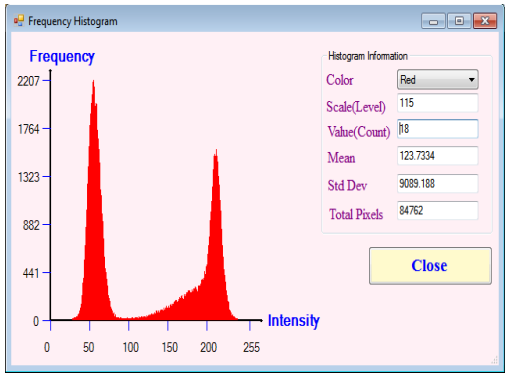

Figure 7: 1st and 2nd derivative of improved edge line detection one- dimension mask.

The system tested over 2000 palm leaf manuscripts images for line count segmentation and varying draw lines are illustrated in figure 7. The proposed system has two analyses, which are carried out for edge segmentation. The first one is to choose background color in the consecutive text area region from palm leaf manuscript images. The second one is to optimize line count for finding edge segmentation. The results of the 1st and 2nd derivative of improved edge line detection enhance by using one-dimensional mask.

5.1 Optimization of line-counts to locate candidate area of palm leaf manuscripts

Line count analysis is carried out for the various background images. In the figure 8, multiple line count analysis is used for segmentation and error % graph lines. However, in order to meet real-time response requirement, the optimization of line counts procedure was elaborated in the following manner. In practice, a few numbers of n (line counts) are needed to be tested. The experimental investigations indicated that, in general, no more than three different values of n needed to be tried for boundary detection. Notice that the blue background color

threshold value for the overlapped pixel amount was empirically determined and proved efficient.

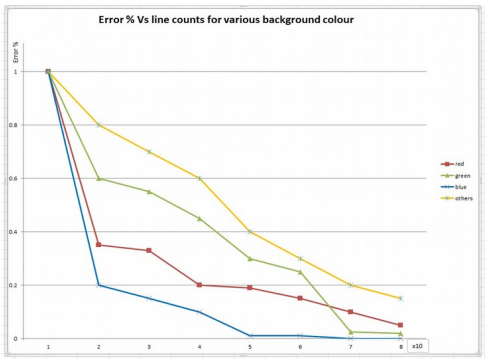

Figure 8: The Error % line counts for various backgrounds graph

The above graph describes to choose optimal line counts for palm leaf object detection. As a result, blue background is the most suitable for image acquisition step in the work of digitizing palm leaf manuscripts. The blue background 50 line counts are enough for this system. In the figure 8, the error % of blue background is the least value than other backgrounds. However, this system is implemented by applying 100 lines for improving accuracy without any errors.

6. Conclusion

According to the research work and results we can conclude as follow:

(a) Fast and robust method for ROI finding can be cropped automatically in palm leaf area images by using one-dimension segmentation.

(b) Line counts analysis is done to optimize or choose suitable amount of lines to find edges.

(c) The empirical method shows the result that in the future what background color should be used in the image acquisition process to digitize palm leaf manuscripts.

(d) The results of this research have been using by Pali-text society to crop palm leaf area automatically.

(e) The output cropped images will be used as the input images or palm leaf manuscripts data set to organize the next text extraction work.

References

[1] E. Jebamalar Leavline and S. Sutha, “Gaussian noise removal in gray scale images using fast Multiscale Directional Filter Banks”, Proceedings of International Conference on Recent Trends in Information Technology, Chennai, India, 2011,June, pp.3-5.

[2] 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.

[3] Nicolas Forcadel, Carole Le Guyader, “Generalized fast marching method: application to image segmentation”, 2008, July, Volume 48, Issue-1 pp.189-211.

[4] Zahra Aareizadeh, Reza P.R., Optik, “A cursive color image edge detection method using Geen’s function approach”, International journal for Light and Electron Optics, 2013, November, Volume 124(21), pp.4847- 4853.

[5]Olga Miljkovic, “Image pre-processing Tool”, Kragujevac J. Math. 32, 2009, pp. 97-107.

[6] Anastasia Rita Widiarti, Agus Harjoko, Marsono, Sri Hartati, “Preprocessing Model of Manuscripts in Javanese Characters”, Journal of Signal and Information Processing, 2014, May, pp. 112-122.

[7] Tzu-Heng Henry Lee, “Edge Detection Analysis”, Graduate Institute of Communication Engineering, National Taiwan University,Taipei, Taiwan, ROC, 2010.

[8] http://www.palitext.com/subpages/thaton.htm

[9] Bindu S moni, G.Raju, “Modified Quadratic Classifier and Directional Features for Handwritten Malayalam Character Recognition Computational Science New Dimensions & Perspectives”, IJCA Journal, 2011

[10] Chacko, B.P.; Babu, A.P, “Online sequential extreme learning machine based handwritten character

recognition”, Students' Technology Symposium, 2011, IEEE.
[11] M. Heath, S. Sarkar, T. Sanocki, K. Bowyer, “A robust visual method for assessing the relative performance of edge-detection algorithms”, IEEE Trans. Pattern Anal. Machine Intell. 19, 1997, December, pp. 1338–1359.
[12] M. Heath, S. Sarkar, T. Sanocki, K. Bowyer, “Comparison of edge detectors: methodology and initial study”, Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 1996, pp. 143–148.