Image Segmentation of Malayalam Palm Leaf Manuscripts

Geena K P
Research Scholar
Kannur University
geena.gopal@gmail.com

G.Raju
Associate Professor Head
Kannur University

***Abstract*— A palm leaf manuscript is one of the earliest forms of written media that has enlightened humanity with various subjects such as medicine, astronomy, mathematics and astrology. There are various texts written on palm leaf manuscripts. Therefore, historical handwritten palm leaf manuscripts are important for people who like to learn about historical documents. The main objectives of the paper are segmentation of characters from Malayalam palm leaf document image. Compare the segmentation results of binary and thinned document image. The better result is produced in thinned document image segmentation. Future development will aim at enhancing the performance of the proposed system more data sets will be used to test the prototype in order to verify the fully automated knowledge and information extraction system for ancient Malayalam manuscripts.**

Keywords*— Malayalam Palm leaf Documents, image segmentation, Otsu's algorithm.

I. INTRODUCTION

Document Image Processing is one of the key application areas of image processing. The rising interest in historical palm leaf document image analysis has created many challenges for researchers. Document Image Segmentation is the act of partitioning a document image into separated regions. These regions should ideally correspond to the image entities such as text blocks and graphical images, which are present in the document image. These entities can then be identified and processed as required by the subsequent steps of Automated Document Conversion [1].Various methods are described for processing Document Image Segmentation. They are Layout Analysis, Geometric Structure Detection/Analysis, Document Analysis, Document Page Decomposition, Layout Segmentation, etc. As the technology is enhancing in day to day life, there is huge amount of increase in the number of documents on the web. There is a need for an application that facilitates the user with an efficient retrieval of the information that is needed [2]. Modern technology has made it possible to produce, process, transmit and store digital images efficiently. Consequently the amount of visual information is increasing at an accelerating rate in many diverse application areas. The large amount of these image data are related to text. The information is stored in the form of digital versions and in document management system. Document image retrieval systems are utilized in many organizations which are using document image databases extensively.

Segmentation is to subdivide an image into its component regions or objects. It should stop when the objects of interest in an application have been isolated. The goal of segmentation is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyse [3]. Segmentation could be used for object recognition, occlusion boundary estimation within motion or stereo systems, image compression, image editing, or image database look-up .Some of the major applications of segmentation are Medical Imaging like locate tumors' and other pathologies, measure tissue volumes, computer-guided surgery, etc. Various other fields where image segmentation is being used: locate objects in satellite images (roads, forests, etc.), face recognition, fingerprint recognition, traffic control systems, brake light detection, machine vision, etc. In this work focus on Malayalam palm leaf document image segmentation. Palm leaf manuscripts are different from other documents. The major problems with them analysis are Poor quality (fragility and deterioration over age),Poor contrast, ghosting noise, holes and spots on the media, narrow spaced lines with overlapping and touching components[4].

This paper is organized as follows: section II describes the review of related work section III describes the features of palm leaf document image, section IV presents the experiment and results, and section V describes the conclusion and future work.

II. REVIEW OF 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. Nowadays, no one can point out which the optimal solution is due to different constraints. In [5], a similarity close measure was used to classify the belonging of the pixels, and then used region growing to get the object. Unfortunately, it required a set of markers, and if there is an unknown image, it is hard to differentiate which part should be segmented. A genetic algorithm adapted the segmentation process to changes in image characteristics caused by variable environmental conditions but it took time learning. In [6], a two-step approach to image segmentation is reported. It was a fully automated model-based image segmentation, and improved active shape models, line-lanes and live-wires, intelligent scissors, core-atoms, active appearance models. However, there were still two problems left. It is strong dependency on a close-to-target initialization, and necessary for manual redesign of segmentation criteria whenever new segmentation problem is encountered. The authors in [7] proposed a graph-based method, the cut ratio is defined following the idea of NP-hard as the ratio of the corresponding sums of two different weights of edges along the cut boundary and models

the mean affinity between the segments separated by the boundary per unit boundary length. It allows efficient iterated region-based segmentation as well as pixel-based segmentation. Moreover, in order to understand an image and recognize the represented objects, it is necessary to locate in the image where the objects are [8].

Some promising research work are reported in palm leaf document images, Olarik Surinta and Rapeeporn Chamchong, 2009, have shown the image segmentation of historical handwriting from palm leaf manuscript, In this paper we presented image enhancement techniques for historical palm leaf manuscript document images, otsu's algorithm using images and finally produces the segmented lines and characters using projection profile analysis [9]. Wafa Bousallana, Abderrazhak Zahour, Adel Alimi, 2008, has introduced a methodology for the separation of foreground or back ground in Arabic historical manuscript using the backlight intensity normalization algorithm, the performance of the algorithm is demonstrated on by real colour manuscripts distorted with show-through effects, uneven background colour and localized spot [10]. Ntogas, Nikolas, VentZas, Dimirios, 2008, have shown the binarization algorithm for historical manuscripts, for binarization algorithm to introduce an innovative procedure for digital image acquisition of historical documents based on image preparation. The results shown that they could extract the palm leaf image as successfully as using high-quality images [11]. Rapeeporn chamchong, Chun che fung, 2009, presented a method for finding a comparing background elimination approach for processing of ancient Thai manuscript on palm leaf, the proposed method also improves the other local adaptive thresholding techniques with a reduction in the noise and recovery of some of the characters [12].2010 (Rapeeporn Chamchong, Chun Che Fung) optimal selection of binarization technique for the processing of ancient palm leaf manuscript [13]. Latest works reported in palm leaf document images are, graph based approach for back ground elimination and Segmentation of the image (2011, Dr.B.P Mallikarjunaswamy, KarunakaraK)[14].Different Binarization and segmentation techniques which are applied in other languages to analyse palm leaf document, but not works are reported in Malayalam palm leaf document [15]. Develop a character segmentation system from ancient palm leaf manuscripts written in Malayalam language by combining several stages, they are Image acquisition, Image enhancement, binarization and automated selection, Text line segmentation and Character segmentation. The problem of segmentation of touching characters which is very common in palm leaf document images is not addressed. In Malayalam if no successful work reported which address the issue of segmentation of touching characters in palm leaf image. Segmentation are some of the challenges faced in Malayalam Manuscript during this step, the characters are touching, Joined handwritten letters. The characters should not be touching each other. Segmentation of touching characters has been one of the toughest jobs in text recognition and this alone has remained one of the toughest areas of research.

III. PALM LEAF DOCUMENT IMAGE

Palm leaf manuscript is one of the oldest medium of writing in India especially in Southern India. It is also the major source for writing and painting in South and Southeast Asian countries including Nepal, Sri Lanka, Burma, Thailand, Indonesia and Cambodia .Though palm leaf writing was practiced since the ancient times its precise origin is still unclear. Agrawal ascertains, "It is difficult to say exactly when the palm-leaf began to be used for writing. There is no extent of palm-leaf manuscripts in India before the 10th century. However, the palm-leaf was definitely in use earlier than this since it's mentioned as a writing material in several literary works and its visual representation can be seen in several sculptures and monuments [16]. Palm leaf manuscripts are different from other documents. The major problems with them analysis are Poor quality (fragility and deterioration over age),Poor contrast, ghosting noise, holes and spots on the media, narrow spaced lines with overlapping and touching components, Unusual, varying shapes, and different styles of characters, which depend on the writer and even the era of writing.

Fig1. An example page from handwritten Malayalam palm leaf document image.

Palm leaf images normally changes to blurred images by the presence of noise, low or high contrast both in the edge area and image area. Pre-processing an image include, removal of noise, edge or boundary enhancement, automatic edge detection, automatic contrast adjustment and segmentation. As multiple noise damages the quality of nature images, improved enhancement technique is required for improving the contrast stretch in palm leaf images [17]. Many methods for historical manuscript image enhancement are driven by the goal of improving human readability while maintaining the original look and feel of the document. The present written document, also palm leaf documents images are usually necessary to difficult to analyses.

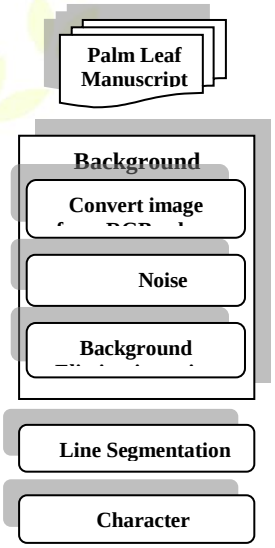


Fig2. Frame work of the proposed system

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. Dataset

For our experiment databases comprising samples of characters from Malayalam palm leaf document image is created. We have collected 7000 Malayalam handwritten palm leaf document images (chadangu Bhasha, keralolpathi, krishnagadaha, Adhyathma Ramayanam, Admananda vivekam, Agnihothra chadangu etc.) from the manuscript library of Malayalam Department, Calicut University. The collected images were scanned at 300dpi. We implement the system to extract data from palm leaf manuscripts. The system processes consist of background elimination, line segmentation and character segmentation.

B. Convert image from RGB color to Gray image

A Color Palm leaf document image is converted to gray image.

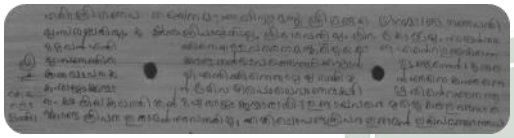


Fig3. An example page from handwritten Malayalam palm Leaf gray image.

C. Background Elimination using Otsu's Algorithm

Otsu's thresholding method involves iterating through all the possible threshold values and calculating a measure of spread for the pixel level each side of the threshold. That is the pixels that either fall in foreground or back ground .the aim is to find the value where the sum of foreground and background spread is at it minimum [13]. To find the optimal threshold (Thr) we can use the following criteria equation which respects Thr.

$$\eta=\sigma 2B/\ \sigma 2Thr$$

Where σ2Thr, that is the total variance. Is independent from the grey level. Only being necessary to minimize the function σ2B, that is the within class variance.

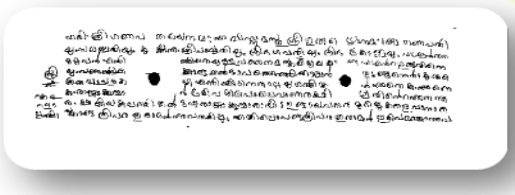


Fig. 4. An example showing a binary image after background elimination.

D. Noise Reduction

The noise, introduced by the optical scanning device or the writing instrument, causes disconnected line segments, bumps and gaps in lines, filled loops, etc. The distortion, including local variations, rounding of corners, dilation, and erosion, is also a problem. Prior to the CR, it is necessary to eliminate these imperfections [17].

E. Character Segmentation

For each image in the database, apply the following method. Find and stored the segmented characters.

Step1: Read color Document image.
Step2: Color image converted to gray image.
Step3: Gray image converted to binary image.
Step4: Remove all objects containing fewer than 30 pixels.
Step5: Show the binary image.
Step6: Binary image change in to thinned image.
Step7: Label connected components in image.
Step8: Measure properties of the image region.
Step9: Plot bounding box in each character image.
Step10: Extracted characters stored in a separate file.

The segmented characters can be used for further analysis of the image. An example segmented handwritten Malayalam palm leaf binary image character and thinned image character is shown in fig.5&fig.6

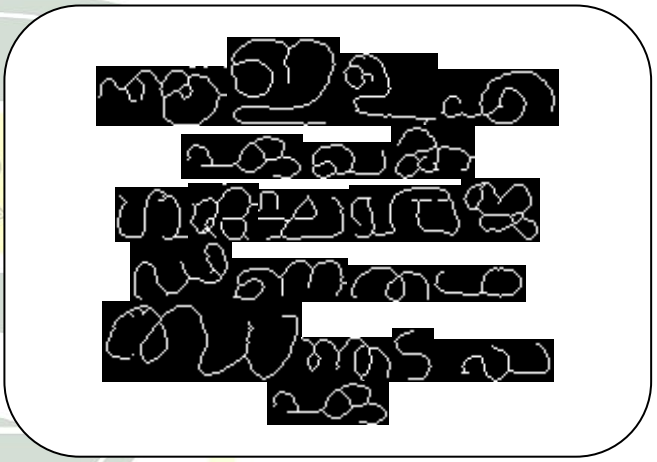


Fig. 5. Segmented binary palm leaf image characters.

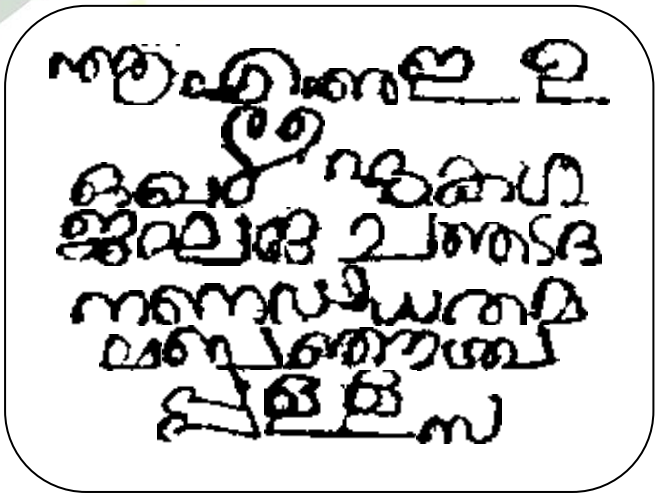
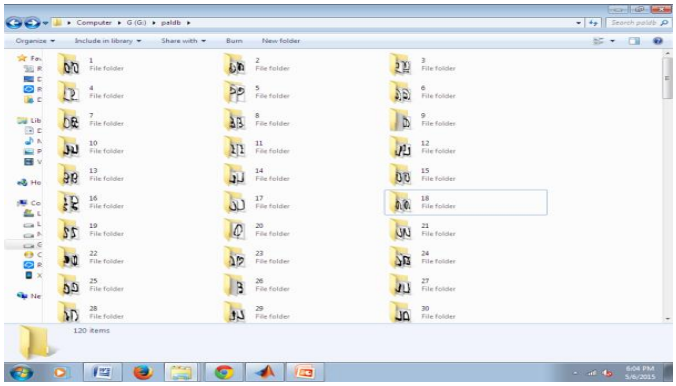


Fig. 6. Segmented Thinned palm leaf image characters.

As a final step, 120 palm leaf segmented characters are stored in a separate file is show in fig.7



CONCLUSION

Future development will aim at enhancing the performance of the proposed system more data sets will be used to test the prototype in order to verify the fully automated knowledge and information extraction system for ancient Malayalam manuscripts. Thinned Palm leaf document image segmentation results are better than binary palm leaf document image segmentation. A framework of optimal selection of binarization techniques will have to be adopted such as evaluation step, feature extraction, feature selection, and ranking of selection Character segmentation has to be improved by considering the touching and overlapping characters.

REFERENCES

[1] G.Nagendhar, D.Rajani, China Venkateswarlu SonagiriV.Sridhar," Text Localization in Video Data Using Discrete Wavelet Transform", International Journal of Innovative Research in Science, Engineering and Technology Vol. 1, Issue 2, December 2012, ISSN: 231–8753.

[2] Nagasudha D, Madhaveelatha and Y Pratap Reddy "Telugu Document Image Segmentation Methods", International Journal of Research and Applications (July- Sep © 2014 Transactions) 1(3): 76-79.

[3] Olarik Surinta and Rapeeporn Chamchong.Image segmentation of historical handwriting from palm leaf manuscript 2008.in IFIP International federation for Information processing, volume 288: intelligent information processing IV. 370-375, 2011.

[4] Itay Bar-Yosef, Itshak Dinstein "line segmentation for degraded handwritten historical document"

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

[6] Chacko, B.P.; Babu, A.P.; Online sequential extreme learning machine based handwritten character recognition. Students' Technology Symposium (TechSym), 2011 IEEE.

[7] Chacko, B.P. Babu, A.P, Pre and Post Processing Approaches in Edge Detection for Character Recognition, Frontiers in Handwriting Recognition (ICFHR), 2010 International Conference on 20 January 2011 .

[8] D.Udaya Kumar, G.V.Sreekumar, U.A.Athvankar .Traditional writing system in Southern India — Palm leaf manuscripts.

[9] A Hybrid Method for Enhancement of Plant Leaf Recognition World of Computer Science and Information Technology Journal (WCSIT) ISSN: 2221-0741 Vol. 1, No. 9,

[10] Wafa Bousallana, Abderrazhak Zahour, Adel Alimi .A methodology for the separation of Foreground/Background in Arabic Historical Manuscript using Hybrid Method.Jounal Universal computer science, vol.14, no.2 (2008).

[11] Ntogas, Nikolas, VentZas, Dimirios. "A binarization algorithm for Historical manuscript, 12th WSEAS International conference on Communications, heraklion, Greece, July 23-25 2008.

[12] Rapeeporn chamchong, Chun che fung).Comparing background elimination approaches for processing of ancient Thai manuscript on palm leaves.Proceeding of the Eighth International conference on machine learning and cybernetics, Baoding, 12-15-july 2009.

[13] Rapeeporn Chamchong, Chun Che Fung, optimal selection of binarization, technique for the processing of ancient palm leaf manuscript 2010.IEEE.

[14] B.P.Mallikarjunaswamy , Karunakara K "Graph Based Approach for Background Elimination and Segmentation of the Image" Research Journal of Computer Systems Engineering- An International Journal, Vol 02, Issue 02, June, 2011.

[15] Geena K.P,Raju G," View Based Feature Extraction and Classification Approach to Malayalam Palm Leaf Document Image" International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 5, October 2014

[16] Zhixin Shi , Srirangaraj Setlur, Venu Govindaraju, "Digital Enhancement of Palm Leaf Manuscript Images using Normalization Techniques", IEEE conference, Computer Vision, Graphics & Image Processing, 2008, P.687-692.

[17] Vikas J Dongre , Vijay H Mankar" Devanagri Document Segmentation using Histogram Approach", "International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.1, No.3,August 2011. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.