5 ICLLCE 2016-12 R.S.Sabeenian

Digitalization of Tamil Characters in Palm Leaf Manuscripts Using Image Processing Technique

R.S.Sabeenian[a*], M.E.Paramasivam[b], P.M.Dinesh[c]
Sona SIPRO, Sona Signal and Image PROcessing Research Center
Sona College of Technology, Salem – 636 005, Tamil Nadu, INDIA.
* Corresponding Author : sabeenian@sonatech.ac.in, sabeenian@gmail.com

ABSTRACT

Four Centuries ago, knowledge transfer from experts to learners was mostly through oral recitation. Script based knowledge transfer was catered by scribing on rocks, metals, cloth and leaves. Due to the abundant availability, the southern part of India mainly used 'Palm Leaves' as the base material for scribing. These ancient methods of writing and protecting manuscripts vanished over a period of time. Today, due to climatic changes and aging factors these manuscripts have started to degrade. To preserve the information present on the palm-leaves, digital images of each leaves have been taken up and stored. We have developed a 'Semi-Automatic System', with primary emphasis on Tamil Manuscripts. In order to have only the textual information for storage and segmentation purposes, it is mandatory that these images are binarized. Binarization of degraded document images has been a challenge to many computer scientists. Binarized characters are segmented and fed to a Character Recognition module capable of replacing the handwritten version to a printed version. The obtained printed version cannot be a fool-proof one, hence a linguistic scholar capable of reading both the handwritten and printed version needs to verify and confirm the final printed version. However, this will drastically reduce the time as compared to the current way of doing things. The development of our module will be of great use to linguistic scholars for quicker scrambling of texts from palm leaves and hence write variety of annotations to a single text.

Keywords: Palm leaf, binarization, character segmentation, CCCMA (Combined Connected Component and Minimal Area)

Introduction

Centuries ago, despite the knowledge of scripts, the most preferred way of knowledge transfer from teacher to student was by Word of Mouth. Even in this digital era, a number of religious teachings in many parts of the world follow this method.

For example, in Tamil Nadu, a state in Southern India, the process of teaching Sanskrit verses from Vedas is termed as *Santhai Solluthal* (சந்தை சொல்லுதல் .(In this method, the teacher spells out a stanza once, followed by the students repeating it twice.

Over the centuries, due to various political and socio-economic changes, there arose a need for alternative approaches of knowledge transfer to generations. Even before the Egyptians invented `*Papyrus'* as a writing material, scholars in India used tree barks, skin hides, rocks, leaves and many other base materials for writing.

Under the leaf category, the commonly used ones were the palm-leaves, due to its wide availability in India. The process of conditioning (Samuel, 1994) palm-leaves for scribing was a day-to-day activity in every household. Scribing refers to the art of writing

on palm-leaf using a sharp metal called Stylus. The oldest available palm-leaf manuscript in India, containing a script of Indian drama dates back to 2 A.D (Diskalkar, 1979).

Scribing practice on palm-leaf manuscripts was prevalent even at elementary schools in Tamil Nadu, until the first Tamil book *Tampiran Vanakkam* (Doctrina Christrum) was printed in 1578. This introduction of printing press made this art of scribing palm-leaves 'uncultured' among scholars. Almost after five centuries from then, today, the linguistic community has only a few people who can read, scribe and interpret information on these manuscripts.

Palm-leaf being organic in nature is susceptible to be attacked by termite and hence is destroyed in a few hundred years. A few samples of very highly degraded palm-leaf manuscripts preserved by Bharathidasan University are shown in Figure 1.

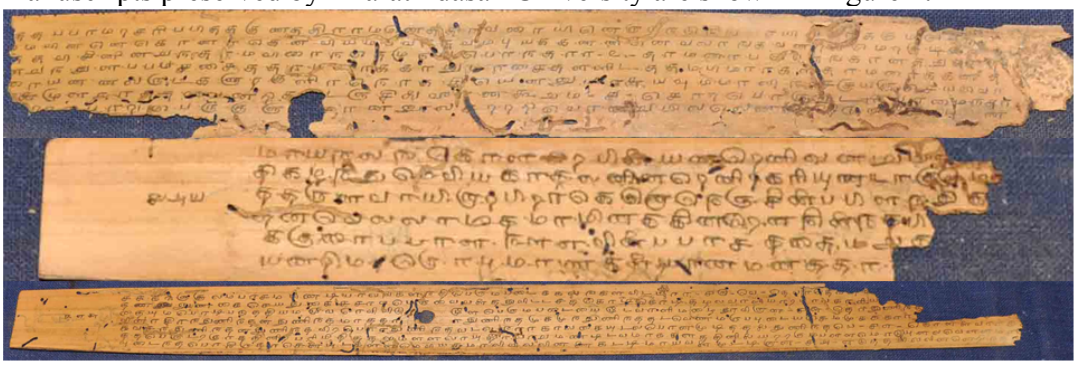

Figure 1.: Highly Damaged Palm-Leaf Manuscripts titled *Kamba Ramayana*

Before the induction of printing press, the method adapted to conserve information present in such degrading leaves was, to immediately arrange for rewriting these contents to a fresh set of leaves. Later, these degraded manuscripts were disposed (Broo) off after rigorous ritually practices.

With the tradition of scribing being forgotten over the years, it was amusing that people had continued the practice of disposing without copying the contents (Samuel, 1994). Just to imagine, it is painstaking to think as how many such manuscripts with most vital literatures would have been lost over the years.

Over years, the tradition of preserving palm-leaf manuscripts has been forgotten due to the large domination of paper. The imaging technology has given a supporting hand to International organizations (UNESCO) (Memory of the World Programme - UNESCO, 2016), Governmental Organizations (National Mission for Manuscript, 2016) and Non-Governmental Organizations (to list a few – (Institute of Asian Studies, Chennai, Tamil Nadu, India, 2016), (Tara Prakashana, 2016), (Chinmaya International Foundation, 2016), (Agama Academy, 2016), etc …), for photographing manuscripts requiring immediate preservation. There are two points described in the following subsections.

Characteristic Nature of Tamil Script

The exact descendant of Tamil characters is from the 7th century *'Vattaeluthu'* script. These *Vattaeluthu* characters had its own complexity, though it was a simplified version of *Grantha* scripts. The transition of *Vattaeluthu* to today's Tamil characters would have happened when scripts had to be written on palm-leaves.

Today, Tamil language has 12 vowels, 23 consonants and two special characters ஸ்ரீ /)sri/) and ஃ/) ah/), thereby forming a character set with 313 characters. The current character set claims its description in '*Tholkappiam'*, a 1st century Tamil Grammar book

and hence claims its classical status today. A complete analysis of current Tamil Character set has been done by (A.G.Ramakrishnan, 2013).

Among the 247 characters in the set, a few of them have a low usage on the palm-leaf manuscripts. Palm leaves being a supple material were endangered of being torn up when a scribe writes on it using a stylus. This constrains prevented writing characters with complex curves and those with dots had to be avoided. Even punctuation marks were evaded due this reason (D. U. Kumar, 2009). A palm-leaf that was damaged during scribing was ignored and once again, the text had to be re-written on a fresh leaf.

The compound character ழ/) za/), considered as the *retroflex approximant* of Tamil, along with its compounded set has been avoided in palm-leaf manuscripts. A sample of characters that have been minimally utilized on palm-leaf manuscripts. டீ /)tii/), ணூ /)Nuu/), தூ/) thuu/), நூ/) Nuu/), லீ/) lii/), லூ/) Luu/) and ஸ்ரீ/) sri/). It is worth to note that the above listed characters have sharp & complex curves and hence were vulnerable for the palm-leaf being injured at the time of scribing itself.

Preserving Palm-Leaf Manuscripts using Imaging Technology

One might think it as a foolish act to look back at these manuscripts living in this digital world. A few Palm-leaf manuscripts contain information related to folk medicine, usage of traditional plants as drugs (termed as Siddha medicine), acupuncture points and other related information.

The technological advancements available today have helped the health-care personnel in identifying and diagnosing many new diseases. Preserving ancient manuscripts shall help health care professionals for looking at alternative medicines for such new diseases. The theme is of interest and under the scope of cultural heritage preservation.

Clearly speaking, this thirst for preservation of manuscripts is not a new theme of the day, but dates back to the 19th century. Dr.U.Ve.Swaminatha Iyer was known (Swaminathaiyer, 2014) for his tireless effort of identifying rare palm-leaf manuscripts and in-turn had them printed on paper.

In the 20th century, to ensure that the human race does not fall ashamed for failing to protect these deteriorating manuscripts, UNESCO (Memory of the World) and many other Governmental (NAMAMI in the Republic of India) and Non-Governmental Organization (Institute of Asian Studies, Taraprakashan, Chinmaya International Foundation, Agama Academy, etc …) started up activities on a war-footing basis.

The two major goals were (a) to improve the life of manuscripts using organic compounds or chemicals and (b) to provide an assured visual reading of manuscripts that have started to degrade. While the first goal involves manually applying chemicals / organic materials, the second goal encompasses photographing of degraded manuscripts.

The inducement to protect ancient manuscripts is not only a requirement in India, but has been a need over the entire globe. Each country has its own set of manuscripts of interest for protection.

For example, the Mahatma Gandhi Institute, Moka, Mauritius has preserved the list of slave workers who had entered the land of Mauritius in the previous century. Lanna manuscripts are available in Thailand. Lanna language is obsolete today and has been replaced with Thai scripts.

Literature Review

The mundane used in the character recognition from Palm-Leaf Manuscripts involves the following: a. Preprocessing (Involving noise removal) b. Binarization c. Segmentation

of Characters d. classifying characters to their corresponding class. A number of available filtering methods can be used for noise removal on the photographed image of palm-leaf manuscripts. There can be no unique method of filtering that can be applied for palm-leaf manuscripts, as the color of the manuscripts shall vary from leaf to leaf.

Binarization of scanned pam leaf manuscripts is a primary pre-processing technique to eliminate the background color and other information not required for character recognition. Sabeenian, Paramasivam and Dinesh (Sabeenian, 2016) have tried to analyze the various binarization techniques developed for document images on palm-leaf manuscripts.

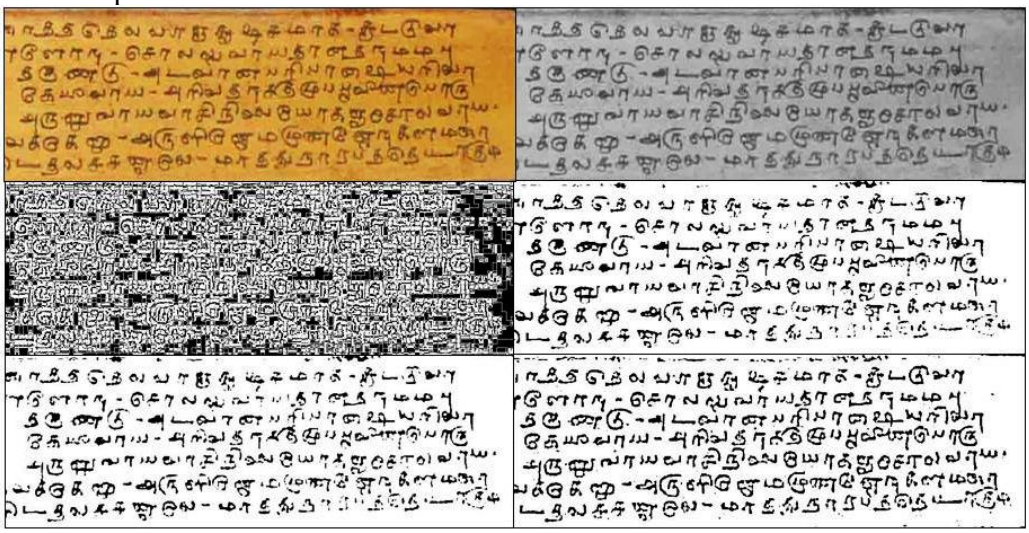

Figure 1: Binarization of Palm Leaf Manuscripts : Results of Sabeenian, Paramasivam and Dinesh (Sabeenian, 2016)

Saxena (Saxena, 2014) has introduced a very novel and effective method of binarization especially for palm-leaf manuscripts. The method proposed by Saxena was successful in improving the readability in many stain affected document images.Chun Che Fung and Rapeeporn Chamchong (Fung, 2015) have carried out an extensive survey on various classical binarization techniques that can suite for palm-leaf manuscripts using SVM.

Jean-Luc Chevillard (Chevillard, 2003) had proposed an encoding system using neural networks for Tamil palm-leaf manuscripts. Chevillard has given a complete Tamil character set utilized in the 18th century in his literature. Almost from then on, very few scientists have taken up this specific area of application research.

Vijaya Lakshmi, Panyam Narahari Sastry and Rajinikanth (T.R. Vijaya Lakshmi, 2016) have recently approached a novel method of character recognition using a contact type 3D profiler. The approach also has eliminated the problem of background removal. Rapeeporn Chamchong and Chun Che Fung have approached the problem of character segmentation from Lanna manuscripts with projection profiles of each character. The work has also focused on touching characters, for which Contour Tracing Algorithm has reported an accuracy of 93.99%.

Olarik Surinta and Rapeeporn Chamchong (Surinta, 2008) have reported a traditional pipeline for character segmentation at manuscripts present in Mahasarakham University under the Palm Leaf Manuscript Preservation Project. The work has tried to evolve line segmentation with an accuracy of 82.5%.

Lanna scripts have characters with many connected characters. Mahasak Ketcham, Worawut Yimyam and Narumol Chumuang (Ketcham, 2016) have utilized Multi-layer Perceptron training for segmenting characters using projection.

Experiments

Proposed Approach

The binarized version of any image shall contain pixel values of either '0' or '1' with the former representing foreground and later indicating the background. The basic idea behind identifying a character is to identify closely linked '0' pixels in a sea of pixels. The closeness of foreground pixels can be identified using 8-Connectivity or 4-Connectivity (Jayaraman, 2010).

Chaining up closely connected identical pixels shall provide the actual regional area covered by each component. This shall act as the skeleton for segmenting each character from the manuscript image.

The maximum and minimum area of each chained pixels are identified. Both these parameters define the total text area along with the area with which the text can be enclosed. Using the calculated values, the smallest rectangle is constructed. The obtained rectangle shall be a 2D vector enclosing individual character of the manuscript image. The Proposed method is called as Combined Connected Component and Minimal Area (CCCMA) based character segmentation

The challenge in the method arises when certain characters are not properly connected. This is mainly due to the fact that the scriber would have not inscribed to the deepest extent, with a motive to save the palm-leaf from being torn up. We have analysed the algorithm on a set of ten different palm-leaf manuscript images and the segmented images were categorized in their respective classes

Dataset

The online repository of Bharathidasan University has been utilized in this paper. Ten images of palm-leaf manuscripts has been selected from the bunch of around 200 manuscript images. The manuscript '*Agathiyar Vaithyam*' was chosen for carrying out experiments. The images being colour in nature, were subjected to the routine of colour-to-gray scale conversion and binarization before subjecting to segmentation algorithms.

Discussion

The proposed method is a Combined Connected Component and Minimal Area (CCCMA)based character segmentation works well on non-connected characters.This is primarily due to the fact that the character has exhibited a maximum text area and hence the proposed algorithm has provided better results.

The structurally confusing character with த is ந has exhibited a comparatively lower segmentation rate. This is primarily due to the fact that the later character has larger area of non-connectivity.

Results

Figure 1. Scanned Palm Leaf Manuscripts

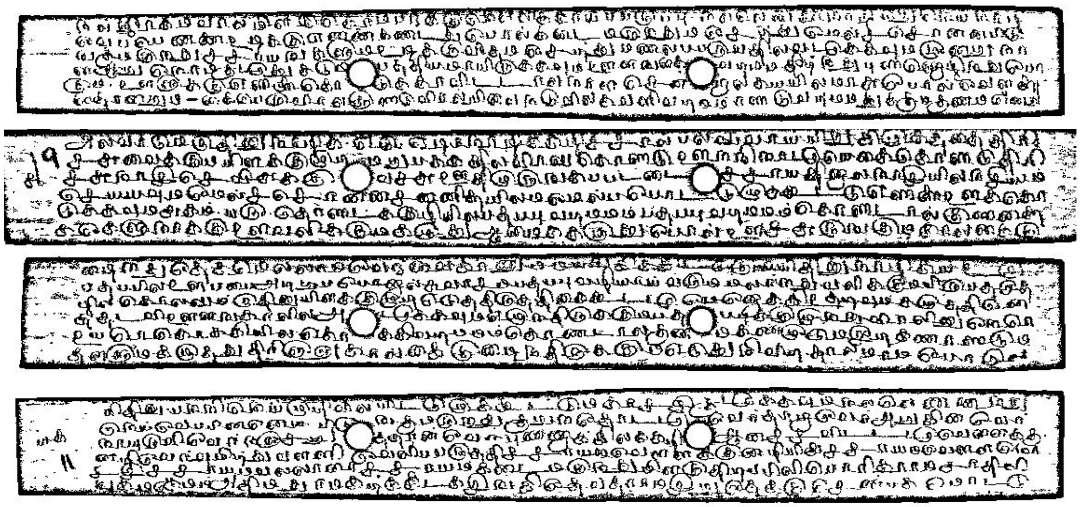

Figure 2. Removing Background using Binarization algorithms

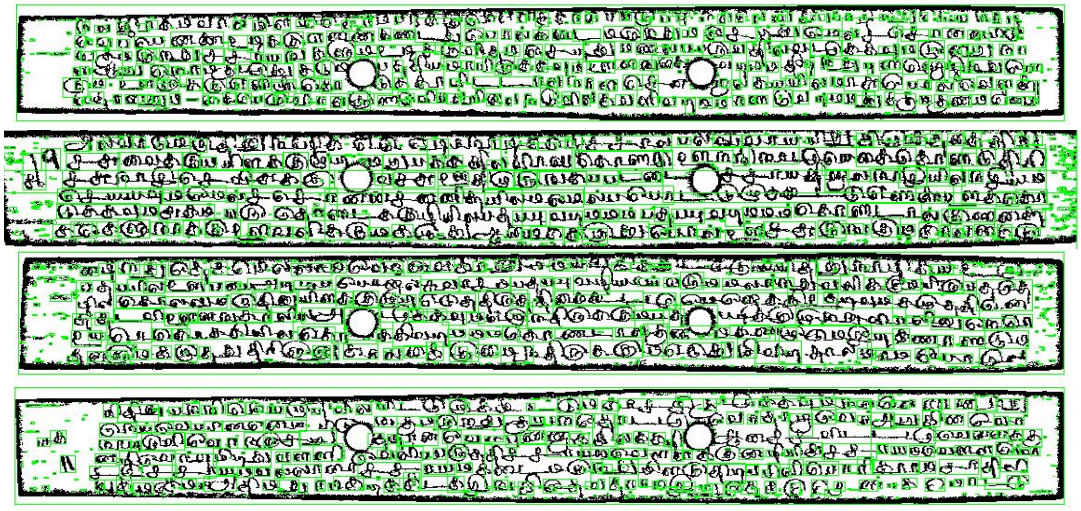

Figure 3. Segmenting Characters

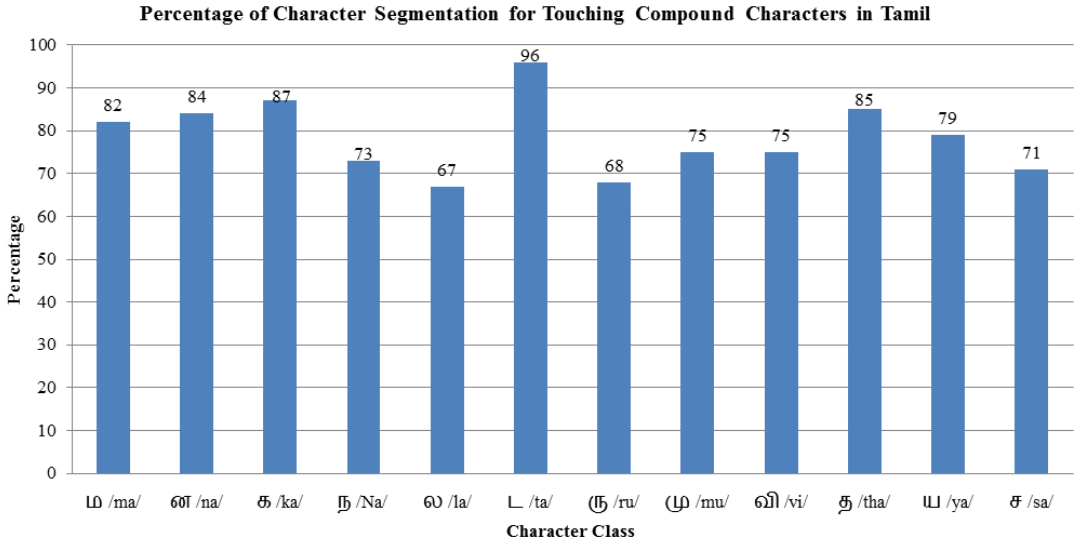

Figure 4. Percentage of Character Segmentation for Touching Compound Characters in Tamil

Conclusion

The proposed method CCCMA works well on non-connected characters. For example the algorithm has exhibited a maximum of 96% for the character 'டை' (/ta/). One connected character for which the segmentation rate is high (85%) is தை (/tha/). This is primarily due to the fact that the character has exhibited a maximum text area and hence the proposed algorithm has provided better results.

The structurally confusing character with தை is ஙை has exhibited a comparatively lower segmentation rate. This is primarily due to the fact that the later character has larger area of non-connectivity.

Linguistic scholar capable of reading both the handwritten and printed version needs to verify and confirm the final printed version. However, this will drastically reduce the time as compared to the current way of doing things. The development of our module will be of great use to linguistic scholars for quicker scrambling of texts from palm leaves and hence write variety of annotations to a single text.

References

World classical tamil conference - a perspective. (2001, August 27). *The Hindu*.

Karunanidhi recalls excerpts of draft on Tamil. (2009, December 28). *The Hindu*.

Agama Academy. (2016, June 1). Retrieved from Agama Academy: http://www.agamaacademy.org/digital-library-en.php

Chinmaya International Foundation. (2016, June 1). Retrieved from Chinmaya International Foundation: http://www.chinfo.org

Institute of Asian Studies, Chennai, Tamil Nadu, India. (2016, June 1). Retrieved from IAS: http://www.instituteofasianstudies.com

Memory of the World Programme - UNESCO. (2016, June 1). Retrieved from UNESCO: http://www.unesco.org/new/en/communication-and-information/flagship-projectactivities/

National Mission for Manuscript. (2016, June 1). Retrieved from NAMAMI: http://www.namami.org/

Project Madurai. (2016, June 1). Retrieved from Project Madurai: http://www.projectmadurai.org/pmworks.html

Tamil Virtual Academy (Erstwhile Tamil Virtual University). (2016, June 1). Retrieved from Tamil Virtual Academy (Erstwhile Tamil Virtual University): http://www.tamilvu.org/library/suvadi/html/index.htm

Tara Prakashana. (2016, June 1). Retrieved from Tara Prakashana: http://www.taraprakashana.org/

Digital Manuscript Gallery, Bharathidasan. (2016, June 1). Retrieved from Digital Manuscript Gallery, Bharathidasan: http://www.bdu.ac.in/suvadi/1.1/

A.G.Ramakrishnan, S. S. (2013). Performance enhancement of online handwritten Tamil symbol recognition with re-evaluation techniques. *Pattern Analysis and Applications*.

Broo, M. (n.d.). *Rites of Burial and Immersion : Hindu Ritual Practices on Disposing of Sacred Texts in Vrindavan.* England: Ashgate Publishing Ltd.,.

Chevillard, J.-L. (2003). A proposal for the digital encoding of palm-leaf tamil manuscripts. *Tamil Internet 2003*, (pp. 109-121).

Cottrell, C. K. (2012). Color-to-grayscale: Does the method matter in image recognition? *PLoS ONE*, e29740.

D. U. Kumar, G. a. (2009). Traditional writing system in Southern India - Palm Leaf Manuscripts. *Design Thoughts (IDC, IIT Bombay)*.

Diskalkar, D. B. (1979). *Materials used for Indian epigraphical records.* Poona: Bhandarkar Oriental Research Institute.

Fung, R. C. (2015). A Framework for the Selection of Binarization Techniques on Palm Leaf Manuscripts Using Support Vector Machine. *Advances in Decision Sciences*, 7 pages.

Jayaraman, E. a. (2010). *Digital Image Processing* . Tata Mc Graw Hill.

Ketcham, M. a. (2016). Segmentation of Overlapping Isan Dhamma Character on Palm Leaf Manuscript's with Neural Network. *Recent Advances in Information and Communication Technology 2016: Proceedings of the 12th International Conference on Computing and Information Technology (IC2IT)* (pp. 55-65). Cham: Springer International Publishing.

Sabeenian, P. D. (2016). Appraisal of Localized Binarization Methods on Tamil Palm-leaf Manuscripts. *WiSPNET.* Chennai: SSN College of Engineering.

Samuel, G. J. (1994). */Uthirum Malargal/ (Tamil).* Chennai: The Institute of Asian Studies.

Samuel, G. J. (1994). */Kumari Muthal Warsaw varai/ (Tamil),.* Chennai: The Institute of Asian Studies.

Saxena, L. P. (2014). An effective binarization method for readability improvement of stain-affected (degraded) palm leaf and other types of manuscripts. *CURRENT SCIENCE*, 489-496.

Surinta, O. a. (2008). Image Segmentation of Historical Handwriting from Palm Leaf Manuscripts. *Intelligent Information Processing IV: 5th IFIP International Conference on Intelligent Information Processing* (pp. 182-189). Beijing, China: Springer US.

Swaminathaiyer, U. V. (2014). */En Sarithiram/.* Chennai: U Ve Sa Noolaga Nilayam.

T.R. Vijaya Lakshmi, P. N. (2016). A novel 3D approach to recognize Telugu palm leaf text. *Engineering Science and Technology, an International Journal*.

APPENDIX

ACKNOWLEDGMENT

The authors would like to thank the All India Council for Technical Education for funding Dr.R.S.Sabeenian with the Career Award for Young Teacher (CAYT) [File No: 11-44/RFID/CAYT/POL-I/2014-15 Dated: 16/03/2015]. The authors owe a lot their Signal and Image PROcessing Team at Sona, the Management of Sona College and their family members for their constant support and encouragement. Special thanks to Dr.John Samuel, Director, Institute of Asian Studies, Chennai, Tamil Nadu, India and Swamy Advayananda, President & Dr.Dilip Kumar Rana, Director, Chinmaya International Foundation (Rashtriya Sodha Sansthan), Ernakulam, Kerala, India for sharing their knowledge on Palm Leaf Manuscripts and script related information. We also thank the Bharathidasan University, Tiruchirapalli for making the digitally stored manuscripts freely available for use.