Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:54 Issue:04:2021
DOI 10.17605/OSF.IO/8DWSQ

TEXT LINE SEGMENTATION IN TAMIL LANGUAGE PALM LEAF MANUSCRIPTS – A NOVEL APPROACH

Dr. M. Mohamed Sathik [1], R. Spurgen Ratheash [2]

[1] Principal, Sadakathullah Appa College, Tirunelveli, Tamil Nadu, India.
Manonmaniam Sundaranar University, Tamil Nadu, India.
[2] Research Scholar, Reg. No: 12334, Sadakathullah Appa College, Tirunelveli, Tamil Nadu, India. Manonmaniam Sundaranar University, Tamil Nadu, India.

Abstract. Segmentation of text lines from palm leaf manuscripts is an essential prior activity for character recognition. The scribers writing style creates intricacy in text line segmentation by low space between text lines and elongated characters placed in the text lines. Inefficient text line segmentation makes unproductive when promoting to character segmentation and character recognition process. The researchers have proposed a new way of text line segmentation algorithm named as Text Line Slicing algorithm for Tamil palm leaf manuscripts. This article explores text line segmentation from the scratch of preprocessing. The identification, segmentation of touching and overlapping text lines by an elongation of the character proves uniqueness of an algorithm. Text Line Slicing provides successful result in Tamil text line segmentation amidst several challenges. This outcome is an evidence of novelty among aplenty of text line segmentation methods in Tamil and other language palm leaf manuscripts.

Keywords: binarization, line segmentation, obstacle, palm leaf, preprocessing, Tamil manuscripts, text line slicing, touching line, overlapping lines.

Introduction

Tamil, one of the most ancient classical languages, has its inscriptions that date back to 600 BC. During the ancient period, many of the literatures, medicinal hints, astrology and much more essential information are present in palm leaves. Lifespan of preserved palm leaf manuscripts is minimum years. The reasons for the dilapidated condition of the manuscripts are weather, fungal and termite. The information of palm leaf manuscripts can be preserved when they are copied into new leaf by the scribers. The palm leaf writing is unique skill that needs patience, practice, and training to the writers. Generally, the Tamil palm leaf manuscripts are written by a pointed needle metal named as stylus [1]. Many of the text lines are not in exact straight line as typed letters. Writing the Tamil characters with stylus creates extension in shapes of the character and makes to touch with the succeeding text lines (Fig. 1). The stylus writing produces the challenges of low space, cross line, touching and overlapping text lines in the process of character recognition from the text images [2]. The successful text line segmentation can lead accuracy when character images are recognizable. In Tamil language, elongation categorizes the strokes as upper part and lower part of the text lines. Tamil character strokes are elongated by nature or in the course of writing by the writers. The letters such as ஞ /thu/, ற/ra/ are the examples of downward and ஙை/nee/, ரு /ree/ are upward elongated characters respectively [3]. The impediment in line segmentation starts from the text lines during the segmentation when they have elongated characters. The proposed Text Line Slicing (TLS) algorithm identifies elongation of the character as an obstacle. The existence or prolongation of an obstacle categorizes the space between the text lines as space without obstacle and space with obstacle. An obstacle touches with the succeeding text lines are considered touching text lines and it

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:54 Issue:04:2021

pervades the text zones of lines in the next text line is known as overlapping text lines. TLS shows an excellence in segmenting text lines in both the ways.

The rest of the paper is organized in section 2 for related works of various text line segmentation methods. In Section 3 discusses about the proposed text line segmentation algorithm. Section 4 presents the results and discussions are presented and the paper is concluded in section 5.

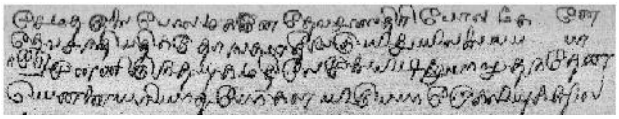


Fig. 1 Tamil palm leaf manuscript

Related works

The touching lines were formed as blocks that can be separated by fixing the bounding box around the components that were present in the lines. If the component spreads over k lines, divide them vertically and fix the cutting locations by using clustering algorithms [4]. In [5], the horizontal projection profile was used to identify the white pixels in the touching text lines and the subsequent histogram was constructed. However, this method detects text line with moderate accuracy. The modified A* Path Planning algorithm has two functions such as intensity difference cost function and vertical distance cost function to construct the segmented line between the touching and overlapping text lines in Khmer palm leaf manuscripts with good recognition accuracy [6]. The minimum horizontal projection values are calculated for each row to identify the touching or overlapping text lines at the end portion of a line and the starting portion of the next line. The minimum values fix the indexing point to segment the overlapping text lines in the printed Tamil scripts [7]. In Adaptive Partial Projection method, the image is divided by vertical columns and the histogram is applied by smoothing to segment the lines [8]. The component break procedure is used to identify the average of the right and left regressions in the connected component to segment the line on handwritten and printed documents [9, 12]. The line adjacency graphs are used to reduce the text components as small called as segments when a set of vertical block runs by Run Length Encoding. The overlapping lines are identified from the connection of vertical runs [10]. The fringe map generated on the binary text-image and hen the peak fringe numbers are located. The filters are applied on the map images and clustering the peak fringe numbers to create the segmentation path between the text lines [11]. An energy map is used to extract the seams from horizontal and vertical orientations and the lower value pixels are removed. The higher value pixel information can be regarded as residing in the text line. The Signed Distance Transform is used to indicate the nearer points and identify the space between the text lines [13].The midpoint detection based method is used to segment the text lines, words, skewed lines, overlapped lines and connected components in handwritten Gurmukhi scripts [14]. An improved piece wise projection based method for handwritten documents to improve an execution speed. The signal approximation using Fourier series and statistical approaches are applied on text lines for segmentation [15]. The Gaussian filter is used to blur the images and discard the pixels which have no local maximum when compare the neighbors. The second order derivatives are used to detect and segment the text lines [16]. A labeling is used to denote the position of an object in a set of features. This tracking method used to segment the text lines in handwritten documents. The method has given failure result in connected components from different lines are close [17]. The text line segmentation based on peaks and valleys in projection profile method yields better result in good spacing text lines. Considering touching lines the result is low in text line segmentation [18]. The starting and ending point of the space can be identified by smoothed horizontal ink

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:54 Issue:04:2021

density histogram in A* path planning algorithm with cost functions to segment the text lines [19].

Proposed method

The text line segmentation starts from the initial process of preprocessing. The preprocessing makes the colour images of palm leaf manuscript in to binary images as it is easier to execute an algorithm to identify an obstacle using image enhancement methods. After preprocessing, the TLS algorithm for text line segmentation is applied on the binary images to analyze the space between the text lines and also categorize the space as space with obstacle and space without obstacle. In the case of space without obstacle, considered as standard images because don't have difficulty in text line segmentation. Space with obstacle categorized as touching and overlapping text lines by an obstacle as in (Fig.2).

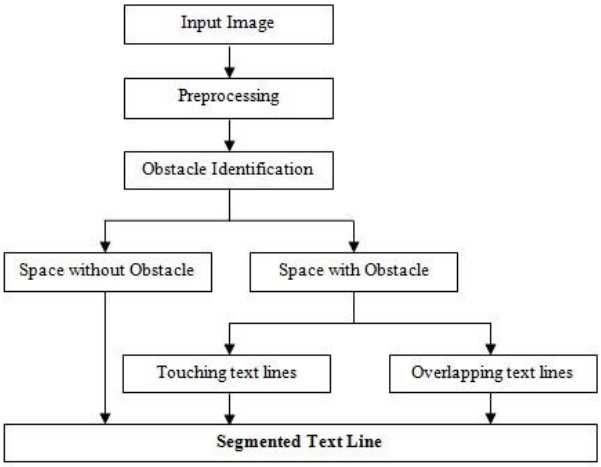


Fig. 2 Architecture of the proposed work

Preprocessing

Digitized palm leaf manuscripts image have dark leaf colour as background and text as foreground. The presence of pickup noise in palm leaf manuscript images when scanned or photo taken by digital camera can be reduced by sharpening and morphology methods. The noise removal is a necessary step to obtain useful information from digital text images. In pre-processing, the background has converted as black and foreground text as white in colors from the range of 0 to 255, into 0 and 1 only then the characters can be clear to process. The background removal and morphology are the methods to promote the images for text line segmentation.

Background removal

Binarization is a process of assigning 0s and 1s using fixed threshold value. The fundamental idea of the fixed binarization method [21] is in the following relation. The background of palm leaf manuscripts can be taken as black by 0 and the foreground text as white by 1. T shows global threshold value 50. The pre-processed images are shows in (Fig. 3).

$$g(x,y) = \begin{cases} 1 \; if \; f(x,y) \geq T \\ 0 \; otherwise \end{cases} \quad (1)$$

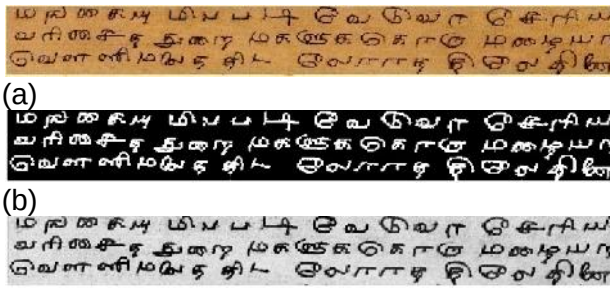
Morphology

In morphological operations, dilation and erosion are the fundamental operations. Addition of pixels with the boundaries of text objects in an image is known as dilation. The reversal of this operation, i.e., extricating the pixels from the text object boundaries is termed as erosion

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:54 Issue:04:2021

[20]. In order to process the text-image in palm leaf manuscripts, the pixels may be added or removed depending on the size and shape of the text. In grayscale morphology, the images are mapped into the Euclidean space or grid E S∪{*r*,-*r*}, the grayscale erosion of the palm leaf image *i* by text boundaries *b* as in the following relation.

$$(i \ominus b)(x) = \bigwedge_{y \in B} [i(x+y) - b(y)] \qquad (2)$$



(a) (b) (c)

Fig. 3 Preprocessed image (a) Input image (b) Background removal (c) Morphology

Text Line Slicing (TLS)

The text line segmentation in Tamil palm leaf manuscripts is a Herculean task and it influences till an end of the character recognition process. An absence of text line segmentation process is not possible for the successful character segmentation and character recognition in Tamil palm leaf manuscripts. The TLS applied on the preprocessed binary palm leaf manuscript images to segment the text lines. The new way of approach in text line segmentation of Tamil palm leaf manuscript images by an obstacle presence between the text lines. Whenever the strokes of the character exceed from the text zone and extend in the space between the lines are known as an obstacle (Fig. 4).

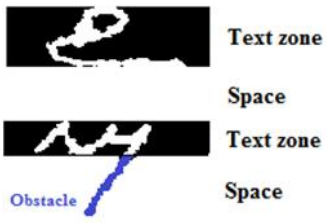


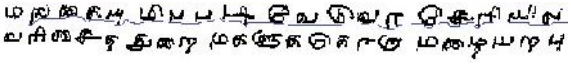
Fig. 4 Obstacle

Space without Obstacle

The TLS, the text lines of Tamil palm leaf manuscripts were the text line has enough space to the subsequent text lines or an elongation of character does not reach the below text line as in (Fig 5) are considered as space without obstacle or standard category. TLS can segment these text lines without any complication.

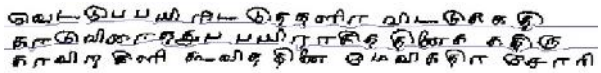
Space with Obstacle

The presence of an obstacle in the space between the text lines can be categorized as two by the length of an obstacle that helps to decide whether touching or overlapping text line. In Tamil character, an



(a)

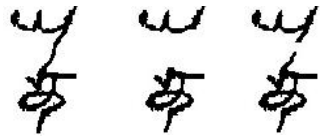
Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:54 Issue:04:2021



(b)

Fig.5 Standard text lines (a) no obstacle (b) obstacle not reached to the below text line.

obstacle is an important part to decide the character. The length of an obstacle extends and reached to the subsequent text line is known as touching text lines as in (Fig. 6). The first line character "உ /you/" is touching with the next line character "த/tha/". In Tamil, the text line segmentation is complicated because if we ignore an obstacle of the character "உ " becomes "ய" and if we cut an obstacle in a fixed length the second line character "த/tha/" becomes "தி/thi/". The overlapping text lines also have the same wrong prediction problem as touching text lines when we precede by existing text line segmentation algorithms. The proposed TLS solves the problem of touching and overlapping text lines by fixing the cutting edge at the end of an obstacle and also prevent wrong predictions of the character.



(a) (b) (c)

Fig.6 Obstacle defined in the textine (a) touching characters (b) ignoring obstacle (c) cut in fixed length

The purpose of text line segmentation is to precede the character segmentation. The touching and overlapping text lines make complicate the text line segmentation and also makes the further process unproductive. An overlapping text line builds complication in text line segmentation. An obstacle pervades the text zone of subsequent lines and mixed up with the character strokes that may precede wrong prediction of the character or different than expected character. The first text line character "ந/na/" extends its elongation up to second text line character "ம/ma/" as in (Fig.7). The proposed line segmentation algorithm TLS segments the character "ந/na/" by fixing the cutting edge in vertically minimum value on the obstacle.

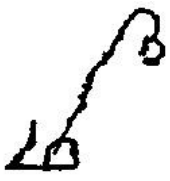


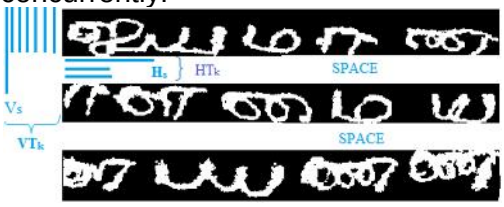
Fig. 7 Overlapping text line

3.2.3 Text Line Slicing algorithm for Tamil text lines segmentation

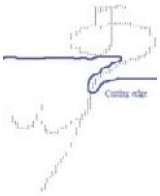
The proposed TLS line segmentation algorithm identifies an extension of character strokes using four variables such as Vertical space (V_s), Horizontal space (H_s), Vertical Track (*VT)* and Horizontal Track (*HT)*. The variable V_s used to count zeros vertically to know the stroke of a character exists in text line of binarized Tamil palm leaf manuscript image. The total columns count of zeros assigned to the variable *VT* and compared with the threshold by the value of 1 denotes that the space has no obstacle and 0for an obstacle. The variable H_s is used to count the zeros in horizontally and the total value assigned to *HT* that

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:54 Issue:04:2021

compare with threshold value. Three values are used to decide whether an obstacle present or not such as zero defines the space between the character in text line; one defines the character has an obstacle and two defines the space not found that means character exists concurrently.



(a)



(b)

Fig. 8 Text Line Slicing algorithm (a) implementation on Tamil palm leaf manuscript (b) cutting edge

The obstacle creates touching text lines can be defined by Connected Component (CC). The connectivity of the character is calculated by the weight of the character using CC. The touching and overlapping text line characters are considered as single character when they are connected to each other as in (Fig. 8). (a) An algorithm implementation proves obstacle identification in the space between the text lines and defines the category of connected characters by connected component and calculates the weight for the character. (b) When the minimum weight of the character identifies the TLS algorithm implements cutting edge to segment the connected text lines. Cutting edge is a breaking point of touching characters in text lines. The CC provides continuation of the character strokes and also vertical stroke values. The minimum value of the character stroke is known as end of an obstacle that has to be fixed as a cutting edge for the text lines.

Results and Discussions

The proposed text line segmentation algorithm TLS proves novelty in text line segmentation in Tamil palm leaf manuscripts as in (Table 1). The touching text lines can be segmented without changing the original shape and preserve all strokes without any loss of information in the characters. In this research, the Tamil language palm leaf manuscripts text lines are categorized by the writing methods of the writer. They are considered as challenges that described by the length of an obstacle and space between the text lines. The segmentation accuracy defines the novelty of an algorithm. Although many systems have been found to recognize the Tamil alphabet, this method has introduced an innovative method of recognizing the Tamil alphabet in palm leaf manuscripts. This has created the process of segmenting the text lines and then segmenting the characters and then recognizing the Tamil characters more accurately. For this research, Tamil palm leaf manuscripts have been taken as 2200 x 300 pixel dimensions for width and height respectively with 300 pixels of resolution. In this section, for the clear vision of challenges, the researchers show the image in a same size of 190 x 280 pixel dimensions for width and height with 100 pixels of resolution. The text line segmentation algorithms are applied on the Tamil palm leaf manuscript images which have all challenges and the results are compared. TLS produced notable results on those challenges than that of other two algorithms.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:54 Issue:04:2021

Table 1 Performance of TLS on Tamil palm leaf manuscripts text lines

	Detection Rate	Segmentation Accuracy	Performance Accuracy
TLS for Standard Images	94.32	98.91	95.96
TLS for Touching Images	92.45	96.58	94.53
TLS for Overlapping Images	90.56	95.84	92.47

CONCLUSION

The text line segmentation is the most important and an initial major process in Optical Character Recognition. The researchers provide a novel approach for Tamil language text line segmentation in palm leaf manuscripts through the proposed algorithm. The touching and overlapping text lines are the major challenges in Tamil palm leaf manuscripts that can be successfully resolved by TLS and it provides an error-free way for other challenges as well. The TLS has advantages from fairly simple to implement, quite fast, and robust for Tamil language palm leaf manuscripts. In future, the TLS can be extended to apply on other language palm leaf manuscripts and Tamil language epigraphs to recognize the characters.

REFERENCES

1. Udaya Kumar, D, Sreekumar, G, V, Athvankar, U, A., “Traditional writing system in Southern India — Palm leaf manuscripts”, IDC.IITB, pp. 1–7, 2009.
2. Ayush Pradhan, Sidharth Behera, and Pushpalata Pujari, "Comparative Study on Recent Text Line Segmentation Methods of Unconstrained Handwritten Scripts", ICECDS, IEEE, pp 3853 - 3858, 2017.
3. Kathirvalavakumar, Thangairulappan, and Karthigaiselvi Mohan, “Efficient Segmentation of printed Tamil Script into characters Using Projection and Structure”, ICIIP, IEEE, pp 484 – 489, 2017.
4. Dona Valy, Michel Verleysen, and Kimheng Sok, “Line Segmentation Approach for Ancient Palm Leaf Manuscripts using Competitive Learning Algorithm”, ICFHR, IEEE, pp.108 – 113, 2016.
5. B. Kiruba, A. Nivethitha and Dr. M. Vimaladevi, “Segmentation of Handwritten Tamil Character from Palm Script using Histogram Approach”, IJIFR, pp.6418-6424, 2017.
6. Dona Valy, Michel Verleysen, and Kimheng Sok, “Line Segmentation for Grayscale Text-images of Khmer Palm Leaf Manuscripts”, IEEE, 2017.
7. Thangairulappan Kathirvalavakumar, and Karthigai Selvi, “Efficient Touching Text Line Segmentation in Tamil Script Using Horizontal Projection”, Springer, pp.279–288, 2013.
8. Rapeeporn Chamchong, Chun Che Fung, “Text Line Extraction Using Adaptive Partial Projection for Palm Leaf Manuscripts from Thailand”, IEEE, pp. 586 – 591, 2012.
9. Himanshu Jain, Archana Praveen Kumar, "A Bottom up Procedure for Text Line Segmentation of Latin Script", IEEE, pp.1182-1187, 2017.
10. QuangNhatVo, GueeSang Lee, "Dense Prediction For Text Line Segmentation in Handwritten Document Images", IEEE, pp. 3264-3268, 2016.
11. Vijaya Kumar Koppula, AtulNegi, "Fringe Map Based Text Line Segmentation of Printed Telugu Document Images", IEEE, pp. 1294-1298, 2011.

12. Ines Ben Messaoud, Hamid Amiri, Haikal El Abed, Volker M¨argner, " A Multilevel Text line Segmentation Framework for Handwritten Historical Documents", IEEE, pp.515-520, 2012.
13. Xi Zhang, Chew Lim Tan, "Text Line Segmentation for Handwritten Documents Using Constrained Seam Carving", IEEE, pp. 98-103, 2014.
14. Rodolfo P. dos Santos, Gabriela S. Clemente, Tsang Ing Ren, and George D.C. Calvalcanti, "Text line segmentation based on morphology and histogram projection", ICDAR, IEEE, pp. 651 – 655, 2009.
15. Vishal Chavan, Kapil Mehrotra, "Text Line Segmentation of Multilingual HandwrittenDocuments Using Fourier Approximation", ICIIP, IEEE, pp. 250-255, 2017.
16. David Aldavert and Marc¸al Rusinol, "Text Line Detection and Segmentation using Second-OrderDerivatives", IAPR International Workshop on Document Analysis Systems Manuscript, IEEE, pp. 293-298, 2018.
17. Setitra, I., Hadjadj, Z., and Meziane, A., "A Tracking Approach for Text Line Segmentation in HandwrittenDocuments", ICPRAM, SCITEPRESS, pp.193-198, 2017.
18. Banumathi. K., Jagadeesh Chandra A. P., "Line and word Segmentation of Kannada HandwrittenText documents using Projection Profile Technique", ICEECCOT, IEEE, pp. 196-201, 2016.
19. Olarik Surinta, Michiel Holtkamp, Faik Karabaa, Jean-Paul van Oosten, Lambert Schomaker and Marco Wiering, "A* Path Planning for Line Segmentation ofHandwritten Documents", ICFHR, IEEE, pp.175-180, 2014.
20. Rodolfo P. dos Santos, Gabriela S. Clemente, Tsang IngRen and George D.C. Calvalcanti, "Text Line Segmentation Based on Morphology and Histogram Projection", ICDAR, IEEE, pp. 651-655, 2009.
21. Nobuyuki Otsu, “A Threshold Selection Method from Gray-Level Histograms”, IEEE, pp. 62 - 66, 1979.