Optimal Character Segmentation for Touching Characters in Tamil Language Palm Leaf Manuscripts using Horver Method

M. Mohamed Sathik, R. Spurgen Ratheash

Abstract: ***An optimality of an automatic character recognition for Tamil palm leaf manuscripts can be achieved only by an efficient segmentation of touching characters. In this article, the touching characters are segmented as a single character to achieve an optimum solution by the recognizer in Optical Character Recognition (OCR). The proposed method provides a novelty in touching character segmentation of Tamil palm leaf manuscripts. An initial process of separation of background image and foreground characters is applied on the palm leaf images by filtering and removing unwanted pieces of characters by noise removal methods. The thickening process overcomes the difficulty of small breakages in the characters. The aspect ratio of the character image can be used to categorize the character such as single or multi touching. Single touching is divided by yet another ways such as horizontal or vertical touching. Finally, the proposed algorithm for* **Horizontal and Vertical** ***character segmentation named as*** **HorVer method** ***is applied on the horizontally and vertically touching characters to segment as independent character. Experimental result produces 91% of an accuracy on segmenting the touching characters in Tamil palm leaf manuscript images collected from various resources and Tamil Heritage Foundation (THF). A novelty method can be achieved in Tamil touching character segmentation by the proposed algorithm.***

Key words : Character segmentation, Pre processing, Touching characters, Tamil character segmentation

I. INTRODUCTION

Tamil language accommodates one of the oldest scripts of the world from 6th century BC. And the birth place of those characters is believed to be Keezhadi, Tamil Nadu, India. The shapes of Tamil characters have evolved from the Greek characters. The script was named as "Tamili" earlier and then called as "Tamil". The writing style of the script is from left to the right. The script has been categorized into two such as vowels (Uyirezhuthu) and consonants (Meiyezhuthu). The vowel has 12 characters and the consonant has 18 characters. The combinations of those two (UyirMeiyezhuthu) categories make 247 characters with one special character '∴' [1].

Ever since the evolution of the language, the characters have specific strokes as its forms. During the period of 17th Century AD, the Christian preacher Constantine Joseph Beschi named who was later known as 'Veeramamunivar' reshaped the forms by means of different strokes and shapes to differentiate the short (Kuril) and long (Nedil) vowels. In 19th Century AD, the great reformer from Tamil Nadu E.V. Ramasamy (Periyar) made changes on the shapes of the characters which continue to date. Prior to the invention of paper, an older civilization of Tamil people used to document their
medicinal hints, information about architecture, astrological ideas, and literatures on palm leaves because it was easily available then in their land. Palm leaf is generally 3.5 cm width and 35cm length as shown in Fig. 1. While writing on palm leaves, the writer holds the palm leaves on the left hand and writes his right hand. The issue here is the way stylus is held makes the difference in the forms of the scripts. Unlike holding a pen with fingers and thumb, a stylus is held while writing within the palm and the four fingers. A stylus is a needle like pen with no ink made of iron rods used to write on palm leaf. The writers write on the palm leaf with minimum pressure without any punctuation so as not to damage the palm leaf. While writing with a stylus, normally the letters or words are written without taking the stylus off the palm leaf and giving room neither for space nor for punctuation. This makes all the difference in the way the scripts are shown as touching characters [2]. OCR only recognizes single characters overlooking the touching characters. Hence, character segmentation is an important phase to provide an optimal solution in recognition phase. The touching characters are divided into two types such as single and multiple touching. The single touching characters are categorized further into two i.e. horizontal and vertical touching as shown in Fig.2. When the characters touch together with the same line characters that is known as horizontal touching as Fig.2a and when the characters touch with the subsequent line characters, they are classified as vertical touching as shown Fig.2b. The characters that touch both ways are called as multiple touching characters as shown in Fig.2c.

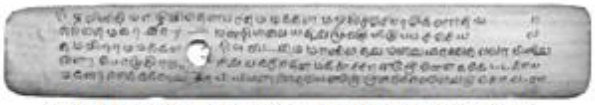
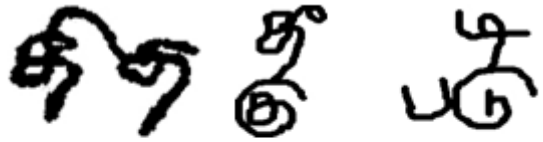


Fig.1 Image of Tamil palm leaf manucript

Revised Manuscript Received on March 30, 2020.

M. Mohamed Sathik, Principal and Research Supervisor mmdsadiq@gmail.com PG and Research Department of Computer Science,
Sadakathullah Appa College, Tirunelveli, Tamilnadu, India. Manonmaniam Sundaranar University, Tamilnadu, India
R. Spurgen Ratheash, Research Scholar, Reg. No: 12334 spurgen@gmail.com Sadakathullah Appa College, Tirunelveli, Tamilnadu, India.



(a) (b) (c)

Fig. 2. Categories of Touching characters (a) Horizontal Touching (b) Vertical Touching (c) Multi Touching

II. LITERATURE REVIEW

The Tamil characters have loops, crossings, junctions and so on and so forth. These highly challenging character segmentation tasks are based on feature extraction using the global and local features. Gabor filter and co-occurrence matrix is used to identify the texture of the character and extracted the global feature. These character segmentation and feature extraction of writings are used by Support Vector Machine (SVM) [3], [4]. The Vertical projection is used to split the words and sub images into individual characters using Structure Based Character Segmentation (SBCS). The Character Threshold is an important part for the separation of the characters while taking an account of width of the character [5]. The characters are segmented horizontally by the template bank. The boundaries are located in edges that are identified by vertical segmentation [6]. The vertical projection profile is calculated for thin images and smoothed using Gaussian low pass filter with standard deviation. The segmentation path can be identified through the projection lines retained that controls an over segmentation of the characters [7]. The multi factorial analysis is used to identify cutting point for the touching characters by the measurement of dissimilarity and aspect ratio with the five fuzzy factors [8]. H. Fujisawa *et al*, introduces a pattern oriented segmentation method to achieve through boundary box to separate the overlapping strokes [9]. The image is known as decomposition by their general features such as dissection method using contextual knowledge by Richard G. Casey *et al* [10]. The word segmentation based on Dominant Overlap Criterion Segmentation performs to separate a stroke group. Besides it and provides the special attention to spatial and temporal features derived from the characteristics and detect the under segmented stroke groups. The SVM used for feedback the stroke group is based on the features to correct the wrong segmentation in the Attention Feedback Segmentation [11]. The touching and broken characters are processed to separate by Graph Partitioning-based Character Segmentation method and it is generalised for multi level writing style of Lamma Dhamma alphabet on palm leaf manuscripts [12]. The character segmentation is categorized by three ways such as 'dissection', 'recognition-based' and 'holistic method'. The dissection methods have predefined rules to obtain segmentation points and it is specifically designed to use contour information [13].

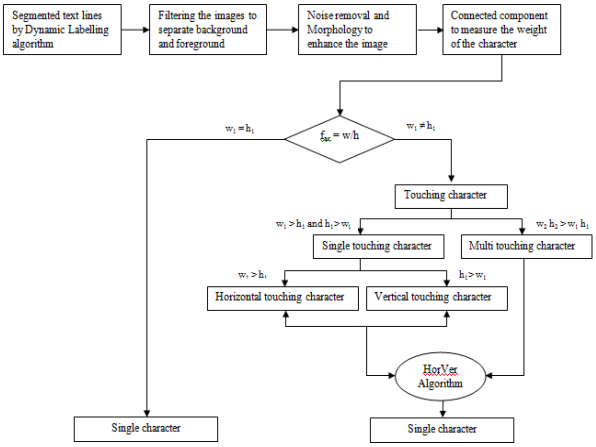


Fig.3 Architecture of HorVer algorithm

The remaining part of the paper has arranged as Section II provides the details of widespread character segmentation methods used in Tamil and various languages by means of a literature survey. The line segmentation method done by the researchers in section III and proposed method of touching character segmentation by HorVer method is explained in section IV. An evaluation results is dealt with in section V with the conclusion in section VI.

III. LINE SEGMENTATION

An OCR in palm leaf manuscripts starts with the primary phase of line segmentation process. An Adaptive Partial Projection (APP) method is used to identify the line numbers and space between the text lines by a piece wise projection method in Thai manuscripts. The second method is known as A* Path Planning (A*PP) is used to identify the touching and partially overlapping characters in text lines by heuristic way. The latter uses by means of various cost functions in Thai and Khmer language manuscripts. In Tamil manuscripts, both the above mentioned methods are ineffective and they change the structure of the character. The Dynamic Labelling Algorithm (DLA) is proposed by the researchers. This method resulted in segmentation of text lines with 96% of Recognition Accuracy (RA). This provides an optimality in Tamil palm leaf manuscripts line segmentation even the characters are deeply touching and overlapping with the proceeding or succeeding line characters [14].

IV. PROPOSED METHOD

The character segmentation is the second major phase in OCR of Tamil palm leaf manuscripts that proceeds after the successful line segmentation process by DLA. The pre-processing stage has image filtering, image sharpening, morphology that are used to remove unwanted particles other than the character present in the lines. Further, they make the character stroke clear. The flow diagram in Fig.3 provides the step by step process of proposed HorVer character segmentation method.

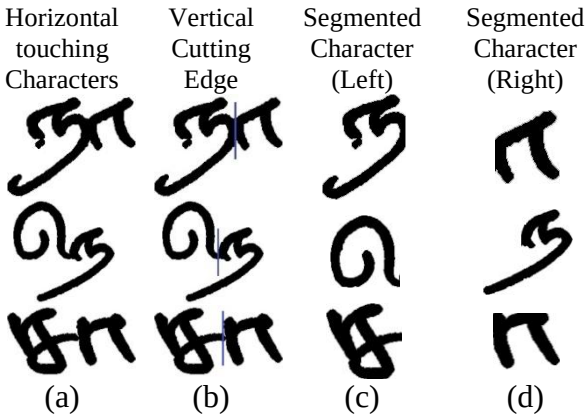


Fig. 4 Segmentation of horizontal touching character

Character Segmentation

In this article the character segmentation process has divided by the following two ways.

1. Identification of Touching Characters

The weight of the character is measured by connected component. A threshold value has fixed to evaluate the weight, when lower value gets than threshold can be considered as single characters and it is automatically separated without any complication. The greater value considered as touching characters using the following factor: *Factor 1: An aspect ratio(ar) of the touching character is larger than automatically separated single characters.*

The above shown factor (f_{ar}) is used to identify the touching characters and that need additional efforts to produce the single characters. The touching characters are defined by

$$\mathbf{f_{ar}} = \boldsymbol{e^a}/(\mathbf{1} + \boldsymbol{e^a}) \quad (1)$$

where $a = w/h$, and w and h is the parameter of width and height of the character. After identification of the touching character, the latter is categorized as horizontal touching, vertical touching and multi touching characters as shown in Fig.2.

Table 1 Possible ways of Single touching horizontal characters

CATEGORY	TYPE	TOUCHING POINT	EXAMPLES
SINGLE TOUCHING (HORIZONTAL)	1		
	2		
	3		
	4		
	5		
	6		

The value of $a=w_1>h_1$ can be identified as a horizontal touching and $a=h_1>w_1$ fall under the category of vertical touching characters. The multi-touching characters have the observation of $a=w_2h_2> w_1h_1$. An analysis of Tamil palm leaf manuscripts gives an observation of maximum possibilities on touching point between two touching characters. That is consolidated in six ways of horizontal touching as in Table 1, four ways of vertical touching as in Table 2, and two ways of multi-touching as listed in Table 3.

Table 2 Possible ways of Single touching vertical characters

CATEGORY	TYPE	TOUCHING POINT	EXAMPLES
SINGLE TOUCHING (VERTICAL)	7		
	8		
	9		
	10		

Table 3 Possible ways of Multi Touching characters

CATEGORY	TYPE	TOUCHING POINT	EXAMPLES
MULTIPLE TOUCHING	11		
	12		

2. Segmentation of Single Touching Character

The segmentation of touching characters is possible when fixing the cutting edge between the characters. The cutting edge is defined as placing the segmentation point to separate two joined characters. Type 1 to 6 in Table 1 shows the possible ways of horizontal touching between two characters. The cutting edge is used for the above shown category characters by calculating their weight in columns. They are identified as the least weight that can predict as the touching point of the characters and also it is considered as cutting edge to segment the character. For this category of horizontal touching characters as shown in Fig. 4a, the cutting edge is vertical way because the weight is calculated by the column of the character as shown in Fig. 4b and the segmented characters are shown in Fig. 4c and Fig. 4d. The vertical touching character as discussed earlier in section 4.1 can be identified by height of the character. The type 7 to 10 in Table 2 shows the possible ways of touching point in vertical touching characters as shown in Fig. 5a. The cutting edge placed between the characters in least weight is calculated by rows. In vertical touching characters, the cutting edge is horizontal as shown in Fig. 5b and the segmented characters are shown in Fig. 5c and Fig. 5d. Type 11 and 12 in Table 3 shows the possibility of multiple touching characters that can be identified by horizontally touching with the same line characters and vertically touching with the subsequent line characters. The process of horizontal and vertical cutting edge provides the touching characters to segment single characters.

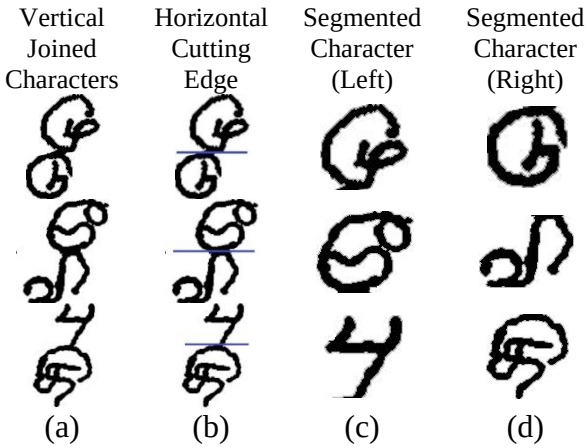


Fig. 5 Segmentation of vertical touching character

V. EXPERIMENTAL AND EVALUATION RESULTS

The HorVer algorithm is applied on line segmented Tamil palm leaf manuscript images by Dynamic Labelling, APP, A*PP algorithms. The proposed method identifies horizontal, vertical and multi touching characters successfully. Besides, it segments the characters as a single character. The experimental values are evaluated by the metrics of DR, RA, and FM based on MM, NN, and o2o. These evaluation criteria and tools are provided by ICDAR 2013 Handwritten Segmentation Contest.

An evaluation of this article by the metrics of one-to-one (oTo) match score, NN and MM [15]. oTo is computed for a region pair based on the evaluators acceptance threshold. Let NN be the total number of ground truth elements and MM be the total number of result elements. An above said three metrics are calculated with the oTo score.

The Detection Rate (DR) is defined by

$$DR = \frac{o2o}{NN} \quad (2)$$

Recognition Accuracy (RA) is

$$RA = \frac{o2o}{MM} \quad (3)$$

and F-measure (FM) is calculated by

$$FM = \frac{2.DR.RA}{DR+RA} \quad (4)$$

VI. CONCLUSION AND FUTURE WORK

The present article sheds light on the character segmentation of touching characters with the category of 'horizontal touching' and 'vertical touching' of Tamil palm leaf manuscript characters. In this article, the comparison of character segmentation method is applied on line segmented images by a novel approach of line segmentation algorithm named as Dynamic Labelling. The proposed DL is invented by the researchers. The latter have used the DL along with two other recent methods of line segmentation such as APP, A*PP algorithms. The HorVer character segmentation algorithm proposed in this article provides optimality about the ins and outs of character segmentation in the Tamil palm leaf manuscripts with 91% of Recognition Accuracy while APP and A*PP provide 87% and 89% respectively from the same value of NN. An overall performance of Recognition Accuracy in character segmentation on line segmented images by APP, A*PP, and Dynamic Labelling listed in Table 4. The Detection Rate of the image and F-Measure shows in Fig. 6 and Fig. 7. The image wise performance of DR and FM is in Fig. 8 and Fig. 9.When the segmented character structure is mismatched with the original structure of the Tamil character, it may solve the mismatch by fixing the least weight in various places on the characters and involve them for continuous iterations in Convolution Neural Network.

ACKNOWLEDGEMENTS

The authors whole heartedly thank Dr. V. Kattalai Kailasam, former Tamil Professor, Tirunelveli, Tamil Nadu, India for the generosity in permitting to access his palm leaf resources. The authors extend their gratitude to Dr. Subashini, Founder of Tamil Heritage Foundation who shared the images of various writing style Tamil Palm Leaf manuscripts

Table 4 Overall performance by HorVer method

ALGORITHMS	NN	MM	O2O	RA
APP	4254	4114	3506	86.15
A*PP	4254	4242	3726	88.78
HorVer	4254	4340	3884	90.63

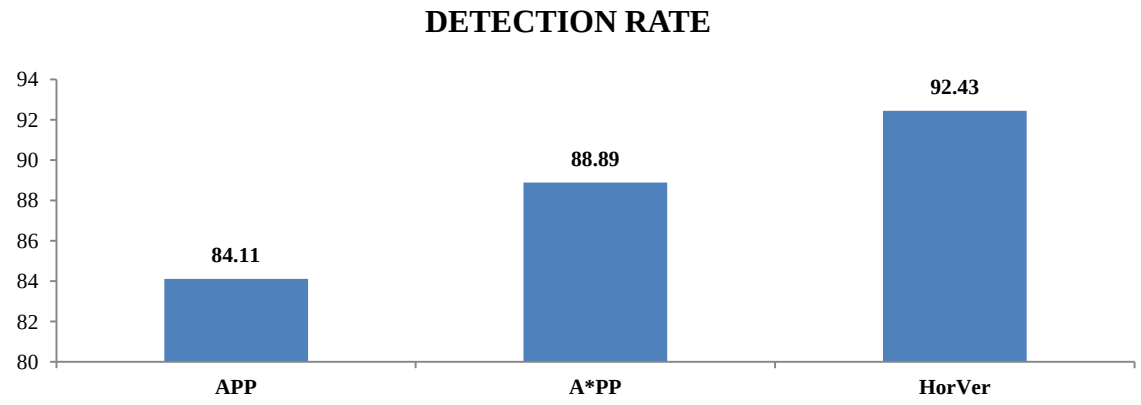


Fig. 6 Comparison of HorVer Detection Rate with APP and A*PP algorithms

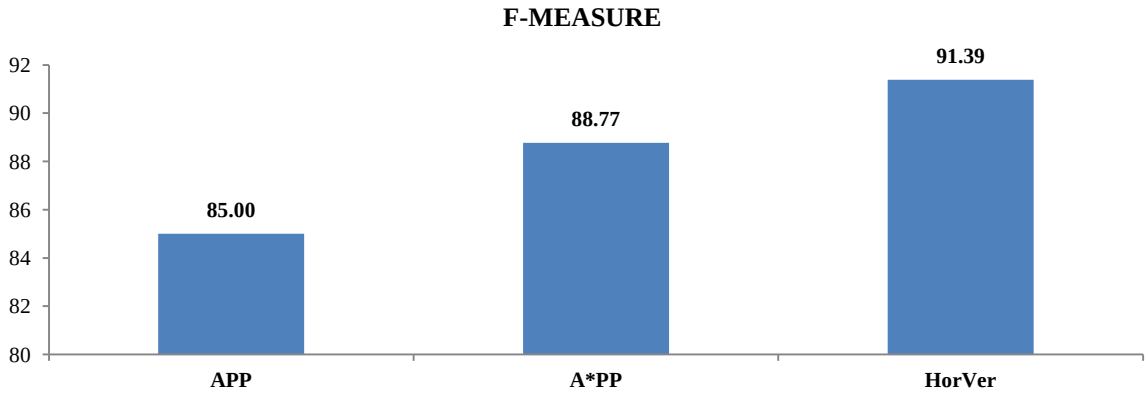


Fig. 7 Comparison of HorVer F-Measure with other two algorithms

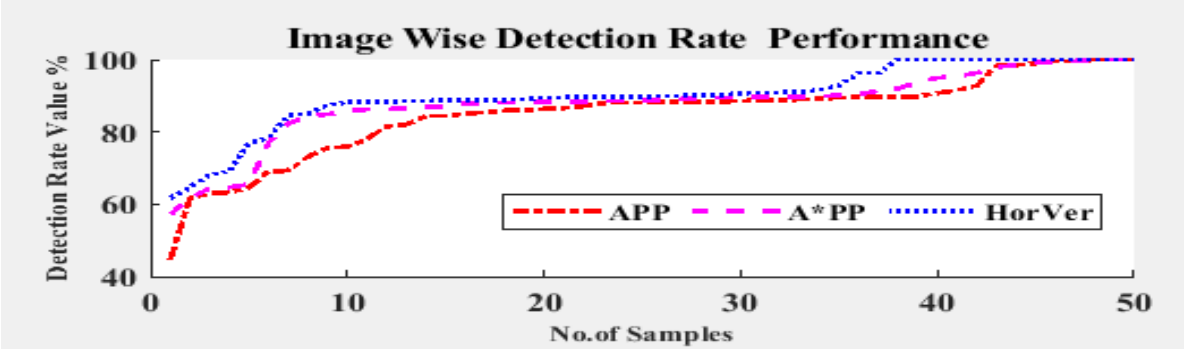


Fig. 8 DR image wise performance of HorVer and other two algorithms

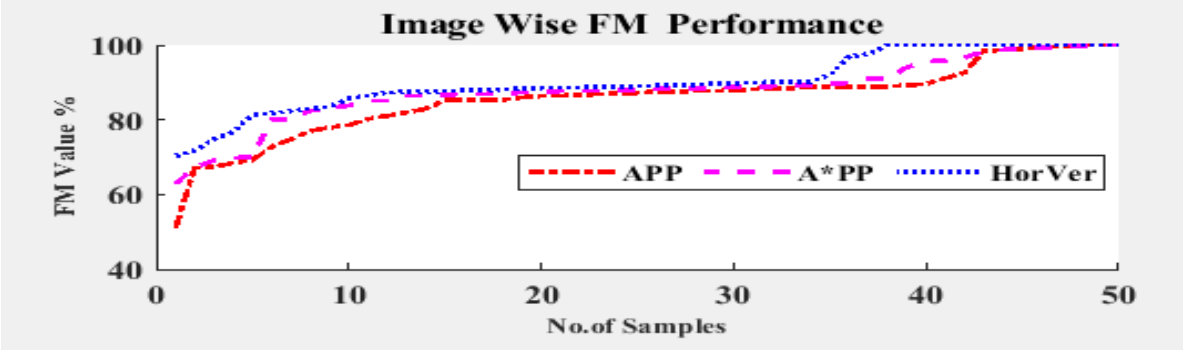


Fig. 9 FM image wise performance of HorVer and other two algorithms

REFERENCES

1. Ponniah.N, Ganesaiyar. C, “Tholkappiyam Eluthadhigara Moolamum” Sunnagam Thirumagal Azhuthagam, pp. 18-19, 1937.
2. D. Udaya Kumar, G.V.Sreekumar, U. A. Athvankar, "Traditional writing system in Southern India — Palm leaf manuscripts", Design Thoughts. July 2009.
3. Thendral T, Vijaya MS, Karpagavalli S, "Analysis of Tamil Character Writings and Identification of Writer Using Support Vector Machine",IEEE IEEE International Conference on Advanced Communications, Control and Computing Technologies, pp 1407 - 1411, 2014.
4. Manigandan T, Dr. V.Vidhya, Dr.Dhanalakshmi V, Nirmala B, "Tamil Character Recognition from Ancient Epigraphical Inscription using OCR and NLP",IEEE, International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), pp 1008 - 1011, 2017.
5. Kathirvalavakumar Thangairulappan, Karthigaiselvi Mohan, "Efficient Segmentation of Printed Tamil Script into Characters Using Projection and Structure", IEEE, Fourth International Conference on Image Information Processing (ICIIP), pp 484 - 489, 2017.
6. J. Tian, R. Wang, G. Wang, J. Liu, and Y. Xia, “A two-stage character segmentation method for Chinese license plate,” Computers and Electrical Engineering, pp 539-553, 2015.
7. Parul Sahare and Sanjay B. Dhok, “Multilingual Character Segmentation and Recognition Schemes for Indian Document Images”, IEEE, pp 1-9, 2018.
8. U. Garain, and B.B. Chaudhuri, “Segmentation of touching characters in printed Devnagari and Bangla scripts using fuzzy multifactorial analysis”, IEEE Proceedings of Sixth International Conference on Document Analysis and Recognition, pp. 805-809, 2002.
9. H. Fujisawa, Y. Nakano, and K. Kurino, “Segmentation methods for character recognition: From segmentation to document structure analysis”, Proceedings of the IEEE, pp 1079–1092, 1992.
10. R. G. Casey and E. Lecolinet, “A survey of methods and strategies in character segmentation”, IEEE Transactions on Pattern Analysis and Machine Intelligence, pp 690 - 706, 1996.
11. Suresh Sundaram, A.G.Ramakrishnan, “Attention Feedback based robust segmentation of online handwritten isolated Tamil words”, ACM Trans. Asian Language Information Processing, pp. 1-25, 2013.
12. Papangkorn Inkeaw, Jakramate Bootkrajang, and Phasit Charoenkwan, “Recognition based character segmentation for multi-level writing style”, Springer, pp. 21-39, Jun 2018.
13. Shi.Z, Govindaraju.V, “Segmentation and recognition 0f connected handwritten numeral strings”, Pattern Recognition, 30(9), pp 1501-1504, Sep.1997.
14. Spurgen Ratheash.R, Mohamed Sathik. M, “Line Segmentation Challenges in Tamil Language Palm Leaf Manuscripts, IJITEE, pp.2363-2367, 2019.
15. R.Spurgen Ratheash, and M. Mohamed Sathik, “A Detailed Survey of Text Line Segmentation Methods in Handwritten Historical Documents and Palm Leaf Manuscripts”, IJCSE, pp 99-103, 2019.

AUTHORS PROFILE

Dr. M. Mohamed Sathik is the Principal of Sadakathullah Appa College, Tirunelveli, India. He received two Ph.Ds majored in Computer Science and Computer Science & Information Technology in Manonmaniam Sundaranar University, Tirunelveli, India. He has many more feathers in his cap by degrees such as M. Tech, MS (Psychology) and MBA. He is pursuing post Doctoral degree in Computer Science. Known for his active involvement in various academic activities, he has attended many national and international seminars, conferences and presented numerous research papers. With publications in many international journals, he has published two books besides having guided more than 40 research scholars. The prolific academician is a member of curriculum development committee of various universities and autonomous colleges in Tamil Nadu, India. His areas of specialization are Virtual Reality, Image Processing and Sensor Networks.

R. Spurgen Ratheash, an Assistant Professor of Information Technology. He is a research scholar at Sadakathullah Appa College, affiliated to Manonmaniam Sundaranar University, Tirunelveli, India. His major research interests include Digital Image Processing, Document Image Analysis and Character Recognition of Tamil language palm leaf manuscripts. He received his MCA degree in Computer Applications from Bishop Heber College, Bharathidasan University, Trichy, India in 2007. He received his MPhil degree in Computer Science and an M.Tech in Information Technology from Manonmaniam Sundaranar University, Tirunelveli, India in 2012 and 2014 respectively.