A Novel Hybrid Model For Tamil Handwritten Character Segmentation

Dr.S.Pannirselvam , S.Ponmani

Abstract— Segmentation is an important task of any Optical Character Recognition (OCR) system. It separates the image text documents into lines, words and characters. The accuracy of OCR system mainly depends on the segmentation algorithm being used. Despite several successful works in OCR all over the world, development of OCR tools in Indian languages is still an ongoing process. Character segmentation plays an important role in character recognition because incorrectly segmented characters are unlikely to be recognized correctly. This paper proposes a segmentation algorithm for segmenting handwritten Tamil scripts into lines, words and characters using Horizontal and vertical profile. The method was tested with different document unconstrained handwritten Tamil scripts, which pays more challenge and difficulty due to the complexity involved in the script. The proposed algorithm results in an efficient extraction of text lines with words and characters by providing average extraction rate and higher segmentation rate.

Index Terms— Handwritten Tamil Document, Pre-processing, Filters, Segmentation

1 INTRODUCTION

Optical character recognition (OCR) refers to a process of generating a character input by optical means, like scanning, for recognition in subsequent stages by which a printed or handwritten text can be converted to a form which a computer can understand and manipulate. A generic character recognition system has different stages like noise removal, skew detection and correction, segmentation, feature extraction and classification. Results of the later stages can affect the performance of the subsequent stages in the OCR process. To make the results of the subsequent stages more accurate, the preprocessing and segmentation play an important role. Most of the Indian scripts are originated from Brahmi script through various transformations. An Optical Character Recognition (OCR) system is the process of transforming human readable and optically sensed data to machine understandable codes. The high performance of any recognition system (OCR systems) depends on the detailed analysis of preprocessing and segmentation operations for removing noises and extracting character components respectively from the input document image. [1]

Segmentation is the process of extracting objects of interest from an image. The first step in segmentation is detecting lines. The subsequent steps are detecting the words in each line and the individual characters in each word. This is a crucial step of OCR systems as it extracts meaningful regions for analysis. This step attempts to decompose the image into classifiable units called character. Segmentation of handwritten text of some Indian languages like Tamil , Malayalam, Kannada, Telugu, Assamese is difficult when compared with Latin based languages because of its structural complexity and increased character set. It contains vowels, consonants and compound characters. Some of the characters may overlap together. Segmentation of words into individual letters has been one of the major problems in handwriting recognition. The complexity involved in the segmentation of characters in the uneven spacing between text lines and adjacent characters. The text lines can also be skewed in some cases.

In the recent past, the number of document images available for Indian languages has grown drastically with the establishment of Digital Library of India. The digital library documents originate from a variety of sources, and vary considerably in their structure, script, font, size, quality, etc. Text line extraction from unconstrained handwritten documents is a challenge because the text lines are often Skewed and the space between lines is not obvious. The complexity involved in the segmentation of the Handwritten Documents for Indian languages like Tamil , Telugu and Malayalam is very well explained in [2]. Curved and non-parallel text lines in handwritten documents also make the segmentation and recognition challenging.

Handwritten text line segmentation approaches can be categorized according to the different strategies used. These strategies are projection based, smearing, grouping, Hough-based, graph-based and Cut Text Minimization (CTM) approach[3].The projection-based algorithm proposed in [4] first obtains an initial set of candidate lines from the piece-wise projection profile of the document .The lines traverse around any obstructing handwritten connected component by associating it to the line above or below. The proposed method is robust to handle skewed documents and touching lines. In smearing based approach technique, consecutive black pixels along the horizontal direction are smeared. If the distance between the white space is within a predefined threshold, it is filled with black pixels. The bounding boxes of the connected components in the smeared image are considered as text lines.

A new approach for text line detection by adopting a state-of-the-art image segmentation technique is proposed in [5]. The authors first convert a binary image to gray scale using a Gaussian window, which enhances text line structures. Text lines are extracted by evolving an initial estimate using the level set method. Grouping approach involves building alignments by aggregating units in a bottom-up approach. Units such as pixels, connected components, or blocks are then joined together to form alignments.

An approach based on perceptual grouping of connected components of black pixels is proposed in [6]. Text lines are iteratively constructed by grouping neighboring connected components based on certain perceptual criteria such as similarity, continuity and proximity. According to the authors the proposed technique cannot be used on degraded or poorly structured documents, such as modern authorial manuscripts.

In this paper a methodology based on projection profile for segmentation of the handwritten Tamil script into lines, words and characters is proposed.

The rest of the paper is organized as follows. Section 2 describes the characteristics of Tamil script, section 3 discusses about the proposed methodology, and section 4 briefly discusses the experimental setup and the results Sections 5 and 6 are performance evaluation and conclusions are discussed respectively.

2. THE CHARACTERISTICS OF TAMIL SCRIPT

Tamil is a South Indian language spoken widely in Tamil Nadu in India. Tamil has the longest unbroken literary tradition amongst the Dravidian languages .Tamil is inherited from Brahmi script. The earliest available text is the Tolkaappiyam, a work describing the language of the classical period. There are several other famous works in Tamil like Kambar Ramayana and Silapathigaram but few supports in Tamil which speaks about the greatness of the language. For example, Thirukural is translated into other languages due to its richness in content. It is a collection of two sentence poems efficiently conveying things in a hidden language called Slaydai in Tamil. Tamil has 12 vowels and 18 consonants. These are combined with each other to yield 216 composite characters and 1 special character (aayuthaezhuthu) counting to a total of (12+18+216+1) 247 characters.

1.2 Vowels

Tamil vowels are called uyireluttu (uyir – life, eluttu – letter). The vowels are classified into short (kuril) and long (five of each type) and two diphthongs, /ai/ and /auk/, and three "shortened" (kuril) vowels. The long (nedil) vowels are about twice as long as the short vowels. The diphthongs are usually pronounced about 1.5 times as long as the short vowels, though most grammatical texts place them with the long vowels.

1.3 Consonants

Tamil consonants are known as meyyeluttu (mey - body,eluttu - letters). The consonants are classified into three categories with six in each category: vallinam - hard, mellinam - soft or Nasal, and itayinam - medium. Unlike most Indian languages, Tamil does not distinguish aspirated and unaspirated consonants. In addition, the voicing of plosives is governed by strict rules in centamiḻ. Plosives are unvoiced if they occur wordinitially or doubled. Elsewhere they are voiced, with a few becoming fricatives intervocalically. Nasals and approximants are always voiced. As commonplace in languages of India Tamil is characterized by its use of more than one type of coronal consonants. Retroflex consonants include the retroflex approximant , which among the Dravidian languages is also found in Malayalam (example Kozhikode), disappeared from Kannada in pronunciation at around 1000 AD (the dedicated letter is still found in Unicode), and was never present in Telugu. Dental and alveolar consonants also contrast with each other, a typically Dravidian trait not found in the neighboring Indo-Aryan languages.

3. PROPOSED METHODOLOGY

In this section segmentation of unconstrained handwritten Tamil script into lines, words and characters is proposed. The proposed method consists of two stages. In the first stage, Preprocessing technique as used to preprocess the image. In the next stage, projection technique is proposed for the segmentation of the text into line, words and characters.

Scanning

A properly printed document is chosen for scanning. It is placed over the scanner. A scanner software is invoked which scans the document. The document is sent to a program that saves it in preferably TIF, JPG or GIF format, so that the image of the document can be obtained when needed. This is the first step in OCR.The size of the input image is as specified by the user and can be of any length but is inherently restricted by the scope of the vision and by the scanner software length.

Phase I : Pre Processing

Pre-processing is a method of eliminating or reducing the noise present in the Image. It consists of various techniques such as binarization, normalization, and another method which is done by various filters. There are efficient filters also available to reduce noise. Image enhancement is the method of improving the quality of the image by increasing contrast, brightness, sharpness etc. The various filter and methods used for pre-processing are discussed in the following sections. The preprocessing stage comprise three steps:

1. Noise Removal
2. Binarization
3. Skew Correction

1. Noise Removal

Noise can cost the efficiency of the character recognition system. Noise may occur due the poor quality of the document or that accumulated whilst scanning, but whatever is the cause of its presence it should be removed before further processing. We have used median filtering for the removal of the noise from the image.

1.1 Filters

Generally filters are used to filter unwanted things or object in a spatial domain or surface. In digital image processing, mostly the images are affected by various noises. The main objectives of the filters are to improve the quality of image by enhancing is to improve interoperability of the information present in the images for human visual.

Median Filter

Median filter is the most prominently used impulse noise removing filter, provides better removal of impulse noise from

corrupted images by replacing the individual pixels of the image as the name suggests by the median
value of the gray level The median of a set of values is such that half of its values in the set are below the median value and half of them are above it and so is the most acceptable value than any other image statistics value for replacing the impulse corrupted pixel of a noisy image for if there is an impulse in the set chosen to determine the median it will strictly lie at the ends of the set and the chance of identifying an impulse as a median to replace the image pixel is very less.

A commonly used non-linear operator is the median, a special type of low-pass filter. The median filter takes an area of an image (3x3, 5x5, 7x7, etc.), sorts out all the pixel values in that area, and replaces the center pixel with the median value. The median filter does not require convolution. (If the neighborhood under consideration contains an even number of pixels, the average of the two middle pixel values is used.) The best known order-statistics filter is the median filter, which replaces the value of a pixel by the median of the gray levels in the neighborhood of that pixel:

y[m,n]=median {x[i,j],(i,j) € w}

where w represents neighborhood centered around location [m,n] in the image.

The original value of the pixel is included in the computation of the median. Median filters are quite popular because, for certain types of random noise they provide excellent noise reduction capabilities, with considerably less blurring than linear smoothing filters of similar size.

2.Binarization

Binarization is a method of transforming a gray scale image into a black and white image through thresholding or Otsu's method be used to perform histogram based thresholding to get binarized image automatically. Otsu's method has been extended for multi level thresholding, called Multi Ostu method. Extraction of foreground (ink) from the background(paper) is called as thresholding. Typically two peaks comprise the histogram gray-scale values of a document image: a high peak analogous to the white background and a smaller peak corresponding to the foreground. Fixing the threshold value is determining the one optimal value between the peaks of gray-scale values [11]. Each value of the threshold is tried and the one that maximizes the criterion is chosen from the two classes regarded as the foreground and back ground points.

3.Skew Detection And Correction

Skew of a document is necessary for many document analysis tasks. Calculating projection profiles ,for example, requires knowledge of the skew angle of the image to a high precision in order to obtain an accurate result. In practical situations, the exact skew angle of a document is rarely known, as scanning errors, different page layouts, or even deliberate skewing of text can result in misalignment. In order to correct this, it is necessary to accurately determine the skew angle of a document image or of a specific region of the image.

Phase II : Segmentation

Segmentation is a process of distinguishing lines, words, and even characters of a hand written or machine-printed document, a crucial step as it extracts the meaningful regions for analysis. There many sophisticated approaches for segmenting the region of interest. Straight-forward, may be the task of segmenting the lines of text in to words and characters for a machine printed documents in contrast to that of handwritten document, which is quiet difficult. Examining the horizontal histogram profile at a smaller range of skew angles can accomplish it. The details of line, word and character segmentation are discussed as follows.

Proposed Technique

After the completion of first stage, the next stage is to extract individual text lines present in the document. In order to extract individual text line, a technique based on projection is used. A projection profile is a histogram giving the number of ON pixels accumulated along parallel lines. Thus a horizontal projection profile is a one-dimensional array where each element denotes the number of ON pixels along a row in the image. Similarly a vertical projection profile gives the column sums. It is easy to see that one can separate lines by looking for minima in horizontal projection profile of the page and then one can separate words by looking at minima in vertical projection profile of a single line. We have used such projection profile based methods for line, word and character segmentation.

Text lines are located using the horizontal projection profile. Then, spacing between lines/words and margins are set to predefined size by means of text padding. Finally, random non-overlapping blocks (of 128x128 pixels) are extracted from the normalized image. Texture analysis is applied to these blocks. First, detect the text lines and empty spaces using the horizontal projection profile(HPP) method (this is simply to demonstrate the uneven lines spacing). Perform a closing procedure on the image using a 3×3 structuring element (only the middle row of the element is set so as to close the image in the horizontal direction to avoid joining text lines). Extract the connected components. Then, compute the minimum, maximum and mean connected component heights.

Proposed Algorithm

Step1: Select the image from the database.
Step 2: Apply median filters to smoothing the image.
Step 3: Binarize the image using Ostu's method
Step4: Apply normalization technique to normalize the image.
Step5: Text lines are located using horizontal and vertical projection profile.

- Segment the lines into words using
  $x_n = \sum_{i=0}^{n} yi, h € [0, v]$
  Where x and y axis represent the horizontal and vertical axis, h represent the height of the image and v represent the size of the image.
- Segment the words into character using
  $x_n = \sum_{i=0}^{n} yi, € [0, b]$

Where x and y axis represent the horizontal and vertical, w represent the width of the image and b represent the size of the image.

Step 6: Segment the word and character Using above step .

Step 7: Repeat step 2 to step 5 for all the images in IDB

4. Experimental Results

The experiments conducted to study the performance of the proposed method. The method has been implemented in MATLAB 7.8. For experimental purpose, we have considered several handwritten document pages collected from different individuals of various professions like school children, undergraduate and postgraduate students, house wives, office employees etc., Our proposed methodology gave an average segmentation rate of 99%, 98.35% and 96% for lines, words and characters respectively.

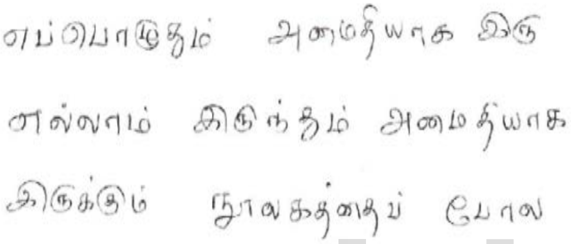


Fig 1 : Original Image

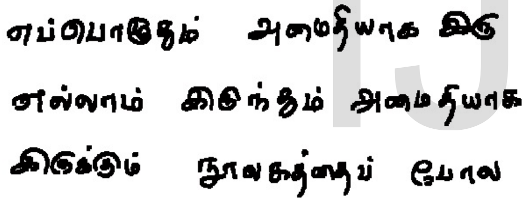


Fig 2 : Pre processed Image

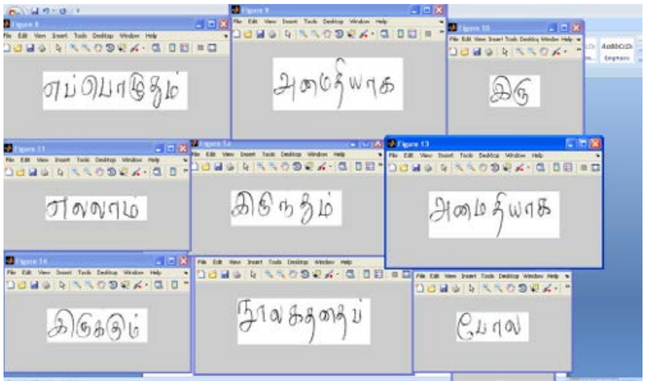


Fig 3 : Word segmented Image

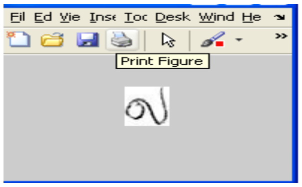


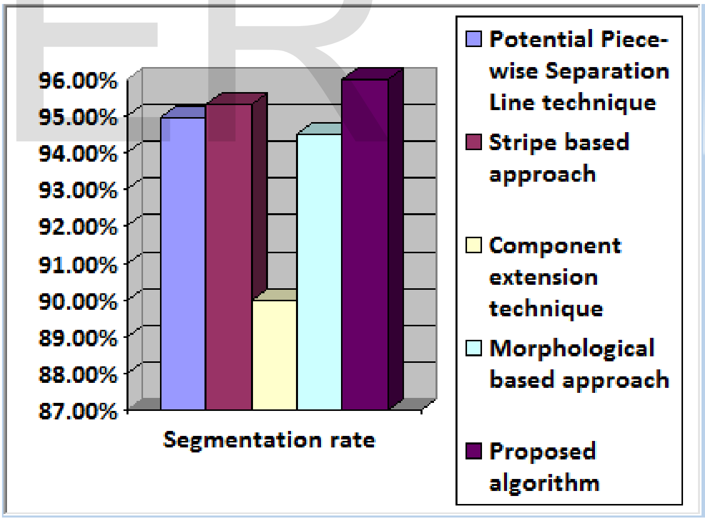
Fig 4 : Segmented character

5. Performance Evaluation

The Table 1 shows the comparison of existing methods with proposed method. To compare our proposed method with the existing work is very difficult as very few works exist in the line segmentation of handwritten tamil document which is experimented on different datasets of complexity. To the best of our knowledge there is no work found in the word and character segmentation for the tamil handwritten documents.

Table 1 . Comparison of proposed method with the existing methods for line segmentation

S.No	Segmentation Method	Segmentation rate
1.	Potential Piece-wise Separation Line technique	94.98%
2.	Stripe based approach	95.32%
3.	Component extension technique	90%
4.	Morphological based approach	94.5%
5.	Proposed Algorithm	96%



Graph 1 Result Comparison Chart

6. Conclusion

In this paper, a segmentation scheme for handwritten Tamil scripts is proposed. The proposed method consists of two stages. In the first stage, Preprocessing technique is used for removing noise and Binarization. In the next stage the projection profile technique is used for segmentation of text into lines, words and characters. The method was tested on totally unconstrained handwritten Tamil scripts, which pays more challenge and difficulty due to the complexity involved in the script. Usage of the proposed algorithm made extracting text lines, words and characters efficiently.

7. ACKNOWLEDGMENTS

Authors would like to thank Dr. S. Pannirselvam, Associate Professor and Head, Department of Computer Science, Erode arts and science college, Erode.who has given the valuable guidance to finish the work, and we would like to thank all writerswho contributed for this dataset.

8. References

[1]. K. Srikanta Murthy, G. Hemantha Kumar, P. Shivakumar and P.R. Ranganath. 2004. Nearest Neighbour Clustering approach for line and character segmentation in epigraphical scripts. In the proceedings of International Conference on Cognitive Systems (ICCS-2004), New Delhi, December 14-15, 2004.

[2]. K.S. Sesh Kumar, A.M. Namboodiri, and C.V. Jawahar.2006.Learning Segmentation of Documents with Complex Scripts. In the proceedings of ICVGIP 2006, LNCS 4338, pp. 749–760, 2006

[3]. ZaidiRazak, KhansaZulkiflee , Mohd Yamani IdnaIdris, EmranMohd Tamil, MohdNoorzaily ,Mohamed Noor, RosliSalleh, MohdYaakob ,ZulkifliMohdYusof and MashkuriYaacob,"Off-line Handwriting Text Line Segmentation : A Review" ,IJCSNS International Journal of Computer Science and Network Security, vol.8 No.7, July 2008 pp 12-20

[4]. M. Arivazhagan, H. Srinivasan, S. N. Srihari.2007. A Statistical Approach to Handwritten Line Segmentation. In Proceedings of SPIE Document Recognition and Retrieval XIV , San Jose, CA, February 2007

[5]. Y. Li, Y. Zheng, D. Doermann, and S. Jaeger.2006. A new algorithm for detecting text line in handwritten documents. In International Workshop on Frontiers in Handwriting Recognition, 2006, pp. 35–40

[6]. L. Likforman-Sulem and C. Faure.1994.Extracting text lines in handwritten documents by perceptual grouping. Advances in handwriting and drawing : a multidisciplinary approach,C. Faure, P. Keuss, G. Lorette and A. Winter Eds, Europia,Paris, 1994, pp. 117-135

[7]. L. Likforman-Sulem, A. Hanimyan and C. Faure.1995. A Hough based algorithm for extracting text lines in handwritten documents.in the proceedings of Third International Conference on Document Analysis and Recognition, Vol. 2, August 1995, pp. 774-777.

[8]. Rajiv Kumar and Amardeep Singh, 2010.Detection and Segmentation of Lines and Words in Gurmukhi Handwritten Text. In the proceedings of IEEE 2nd International Advance Computing Conference,2010,pp 353-356

[9]. G. Louloudis , N. Stamatopoulos , B. Gatos .2009.A Novel Two Stage Evaluation Methodology for Word Segmentation Techniques. In the proceedings of 10th International Conference on Document Analysis and Recognition,2009, pp 686-690

[10]. Richard G. Casey and Eric Lecolinet.1996."A survey of Methods and Strategies in Character Segmentation", IEEE Transactions On Pattern Analysis And Machine Intelligence, vol. 18, no. 7, July 1996 pp 690-706

[11]. R. Plamondon and S. Srihari. On-line and Offline Handwriting Recognition: A Comprehensive Survey. IEEE Trans. on Pattern Analysis and Machine Intelligence, 22(1):63–84, 2000.

[12]. AlirezaAlaei, P. Nagabhushan and Umapada Pal.2011.A Benchmark Kannada Handwritten Document Dataset and its Segmentation. In the proceedings of International Conference on Document Analysis and Recognition, 2011,pp 141-145.

[13]. B. Gatos, N. Stamatopoulos and G. Louloudis.2009."ICDAR 2009 Handwriting Segmentation Contest," Proc. of 10th ICDAR, 2009, pp. 1393–1397.

[14]. A. Alaei, U. Pal and P. Nagabhushan.2011."A new scheme for unconstrained handwritten text-line segmentation", Pattern Recognition, 44 (4), 2011, pp. 917–928.

[15]. V. N. ManjunathAradhya and C. Naveena.2011.Text Line Segmentation of Unconstrained Handwritten Kannada Script.In the proceedings of ICCCS'11, 2011, pp 231-234.

[16]. Mamatha H R and Srikantamurthy K. Morphological Operations and Projection Profiles based Segmentation of Handwritten Kannada Document ,*International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868 Foundation of Computer Science FCS, New York, USA Volume 4– No.5,October 2012*

[17]. R.JAGADEESH KANNAN and R.PRABHAKAR, Off-Line Cursive Handwritten Tamil Character Recognition,WSEAS TRANSACTIONS on SIGNAL PROCESSING, Issue 6, Volume 4, June 2008