

Segmentation and Recognition of Touching Characters In Machine Printed Telugu Documents Using Average Character Widths and Central Moments Features

N. Shobha Rani^{#1}, Pradeep C. H.^{*2}, Sharan J^{#3}

[#] Department of Computer Science, Amrita Vishwa Vidyapeetham
Mysore Campus, Mysore, Karnataka, India

¹ n.shoba1985@gmail.com

² pradeeppanayal@gmail.com

³ sharan2912@gmail.com

Abstract

Accurate recognition of machine printed Telugu documents is one of the principal requirements for the furtherance of the Telugu optical character recognition system. The definite recognition of characters in south Indian languages like Telugu can be realized only when a systematic and ordered segmentation is performed in the segmentation stage of optical character recognition. This paper proposes a recursive segmentation and recognition approach based on the average character widths for the segmentation of touching characters in machine printed Telugu documents like newspapers and text books. The algorithm functioning is based upon the database of trained features of various character components in the Telugu script. The central moment features of the segmented components are used to create a database of various character components in Telugu script. The algorithm had obtained the encouraging outputs in the segmentation process and had achieved an overall recognition rate of around 93-97% in most of the documents experimented.

Keyword: touching character segmentation, average character widths, central moments, Telugu machine printed documents, character recognition.

Introduction

OCR (Optical Character Recognition) refers to a branch of computer science which involves reading text document images and translating images into a form that computer can manipulate. An OCR enables you to convert a text document image directly to a editable file like word processor. OCR plays a very important role in the society as its applications are extensive in the automation of historical scriptures and many other data processing tasks.

Even after numerous experimentations are reported in the area of optical character recognition for South Indian languages [1], the practicality of its applications in automation of textual images into editable documents is very limited. The practical short comings of most of the south Indian OCR's are its poor recognition rates [2]. The barrier for the accurate character recognition is the segmentation stage of OCR. For Roman languages like English, the segmentation process is straight forward, because of the non-existence of structural characteristics that persists in most of the South Indian languages. The South languages like Telugu and Kannada script possess complex structural characteristics like super script, sub scripts along with base character, which are the obstructing factors that complicates the process of segmentation in the OCR. Telugu is considered as the native language of the state Andhra Pradesh and which is spoken by more than 50 million people in South India. Telugu literature is considered to be very prominent in the literature. Many document communications in various Govt. or private organizations are carried out in its native language. In order to perform the automatic processing of all these documents a well standardized and efficient Telugu OCR is very much required. Although many online OCR's for various languages exists, it is not adaptable for all types of documents and more over existing OCR technology will functions fine with many of the constraints like characters should be isolated from one to another and there should no overlapping or partially touching characters in the document image etc. Along with all these constraints, there are even some restrictions like font sizes and font styles, machine printed characters, hand printed characters etc. Even though the recognition of printed documents with all the above mentioned constraints and restrictions has achieved an overall accuracy of around 97-99% [4], the approaches can't be easily extended towards the documents containing touching characters or overlapping characters of various fonts styles and sizes. The approaches like connected component analysis is efficient in performing segmentation of overlapping characters that are not touching to one another and projection profile analysis is not a suitable method of segmenting overlapping or touching lines in the Telugu text because the obstructing nature of super scripts and sub scripts in the spaces between the lines. Some of the heuristic or hybrid approaches like the multiple projection profiles, recursive XY-cut approach and water reservoir principle are devised to segment the overlapping and touching characters these are applicable to segment the numerals, strings and some of the touching fonts with respect to languages like Tamil, Gujarati and other scripts. The segmentation of languages like Telugu requires a hybrid approach that can identify touching components, overlapping components and other compound characters with regard to independence from variety of font styles and sizes, since segmentation of document image plays a crucial role towards achieving the highest accuracies in the process of classification and recognition.

All the above factors had motivated us to devise a hybrid approach that can perform the segmentation of touching and overlapping characters in printed Telugu documents through which expected recognition accuracies can be reached.

Literature Survey

In the perspective of improvising the results of segmentation and recognition many of the researches are carried out from 1970's. Some of the approaches devised during these researches are reviewed as follows.

Abdul Rahiman et. Al [1] had given an overview of variety of characteristics of scripts, recognition and segmentation techniques that could be employed for south Indian languages. Rinki Singh et. Al [2] proposed an approach of classifying Telugu script based upon features like character height, character width, number of horizontal lines (long and short), the number of vertical lines (long and short), number of slope lines, special dots. The extracted features are passed to neural network where the characters are classified by supervised learning of Back Propagation algorithm. The approach devised is suitable for segmentation of isolated Telugu characters in the document. Mamatha H. R et. al [3] has proposed an scheme for recognition of handwritten Kannada numerals using run length count and directional chain code features and had attained an accuracy of around 96%. The method works with reduced features for recognition of numerals. Mamatha Hosalli Ramappa et. al [4] has proposed an approach of segmentation of handwritten Kannada documents using bounding boxes approach and obtained an accuracy of over 91% for lines and 70% for words segmentation. J. Bharathi et. al [5] had propped an algorithm for segmentation of touching consonant conjuncts in printed Telugu script using minimum area bounding boxes and achieved an accuracy of 96%. The method uses structural characteristics to identify the location of touching characters which are touching at the junction of bottom and middle zone. M. Swamy das et. al. [6] has devised a method for segmentation of overlapping text lines and characters in Telugu document images using projection profiles, connected components and spatial vertical relationships and nearest neighborhood method is used to cluster the connected components. The results obtained in case of overlapping line segmentation is 100% and around 98% of accuracy in case of overlapping characters. Srinivasa Rao A V et. al. [7] has proposed an algorithm for segmentation of touching Telugu characters under noisy environment using drop fall algorithm. The algorithm is tested on set of touching characters that are regularly occurred in the handwritten documents in offices, schools etc. The accuracy mainly depends upon the noise that is added to the image and fails to handle if broken characters are present. Mamatha H R et. al [8] has proposed an approach of segmenting handwritten Kannada document using morphological operations and projection profiles and attained a recognition rate of 82% for words and 73% for characters. Srinivasa Rao A V et. al. [9] had propped a method of segmenting touching handwritten Telugu characters using drop fall algorithm that uses a moving marble to locate the segmenting locations in the touching characters. The method is based on top and bottom based profiles of characters and fails to segment from left or right portions of characters. Nallapa Reddy Priyanka et. al. [10] had proposed an approach of line and word segmentation for printed Telugu documents using modified histogram and run length based smearing techniques. The method also proved to be generic for various languages like Kannada, Devanagari etc.

It is been evident from the various paper surveyed that, the experimentations on touching character segmentation of south Indian languages is less compared to Roman

languages. Even though some of the experimentations focus on touching character segmentation the accuracies attained during the testing with other types of font styles are not much encouraging since they are restricted to work with noisy environments and datasets created using paint brush and side profiles of touching characters etc. Thus it is a prime necessity to devise an effective algorithm for the segmentation of touching characters in printed Telugu document images.

Proposed Methodology

The proposed methodology is devised into three stages. The first stage encompasses the pre-processing of the document image through which a well enhanced binarized and thinned image is obtained as output. The well enhanced binarized image considered as input to the second stage, where in the document is segmented into lines by applying an approach of connected component analysis [6]. Each of the line segmented is driven as input to the stage three in which a recursive segmentation and recognition approach based upon the average character width and a database of moment features are used to perform the touching character segmentation.

The stage one of the proposed methodology is very much similar to any of the image processing pre-processing methodology; the input document image is an image with noise which consists of some broken characters and highly dilated characters due to the environment where the documents are actually printed. The image will be binarized by using Otsu's thresholding algorithm [7] and then Gaussian filters [7] are applied to in order to eliminate the noise from a binarized image, since there will be some uncertainties in the selection of the exact threshold for binarization. The filtered image will further undergo morphological operations like dilation and erosion [8] to smoothen the image such that the broken characters and highly dilated characters can be rectified in this process.

The pre-processed image from stage one is further segmented into lines in stage two. The compound or connected characters are generated in each line by the adding one or more consonant or vowel modifiers to the base characters. The presence of compound or connected characters in the Telugu script forbids the clear segmentation boundaries from one line to another line. In order to determine the height of a line, the horizontal histogram of the document is analysed only with respect to first 10% of columns of the document beginning with starting character column in a row. Since the document is a printed document where in the document may maintain the equal line heights and equal spaces between the lines throughout the whole document. Determine the starting of every line by analysing the horizontal histogram computed using the first 10% of the columns in the document using which the height 'n' of the line is computed. The starting point of a line is nothing but a row of black pixel count zero followed by its immediate row's black pixel count as nonzero. From the starting of every such row consider the portion of the document which is equal to the line height 'n' of the document image. From the last row of each such chunk considered, perform the connected component analysis with respect to threshold of 8-connectivity among the pixels towards either of the right, diagonally top or bottom direction in order to examine the path of the line. The connected component analysis is performed

by examining each white pixel and its neighbours in the order of right, top right diagonal and bottom right diagonal with respect to 8-connectivity relationships in each direction. This process will be continued till all the starting points of the rows in the document image are exhausted. The figure 1 and figure 2 represents the original input image and the output of segmentation of overlapping lines using connected component analysis.

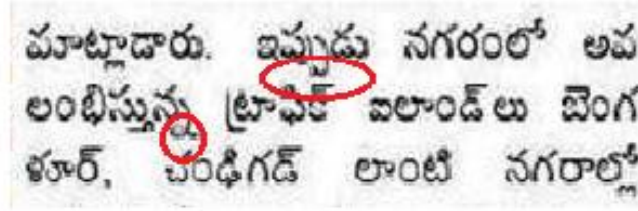


Figure 1: Original Image, Circles Represents The Overlapping Portions Between Lines

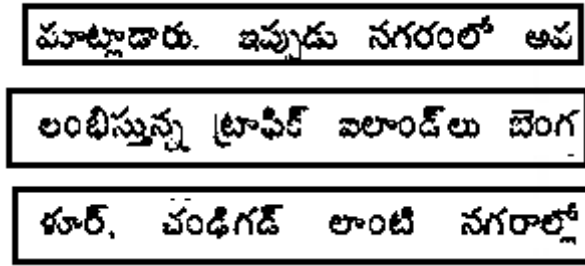


Figure 2: Lines Extracted From A Filtered Image Using Connected Component Analysis

A. Analysis of average character widths:

The character width represents the total number of columns in the image on which a character is super imposed. An average character width is the average of a group of character widths that are closer in distance to each other. The character width of characters in Telugu script or scripts of any South Indian languages like Kannada, Tamil etc varies from one to another. Initially the widths of different types of characters with respect to some of the touching font styles of standard font sizes used in newspaper like images and some other documents related to printed text books are examined to evaluate the average variable widths A_1 , A_2 and A_3 . The width of each isolated character like La, Lu, Ya, Va, U and certain compound characters Sree, Ryu etc are used to determine variable widths A_1 , A_2 and A_3 . The widths of different characters are averaged to only three different average character widths based upon the similarities and proximities among various computed widths.

Let $n_1, n_2, n_3, \dots, n_l$ are the character widths computed from 'n' variable number of isolated characters, then average character width A_k is nothing but the average of a

set of closest variable widths n_j where $j=1, 2, 3 \dots m$. The average character width A_k is computed empirically as,

$$A_k = \sum_{j=1}^m \frac{n_j}{m}$$

Where $n_1, n_2, n_3 \dots n_m$ are the set of character widths which will vary only in terms of one or two pixels column wise and $k=1, 2, 3$ as per the proposed methodology.

B. A Recursive Segmentation and Recognition Approach:

The stage three of the proposed methodology performs the word level character segmentation of touching characters. Each segmented line is further segmented into word using vertical projection histograms. Each word is examined from its right most column or last column; an average character width ' A_k ' is extracted from its last column towards the left direction of the word where $k=1, 2, 3$. i.e., if average character width ' A_1 ' is equal to 10 then from n th column to $(n-10)$ th column the portion of the word is extracted and its moment features are compared with the existing set of features available in the database, if a match is encountered then the segmentation boundary between last character and its preceding character is marked as valid. If a match is not encountered further average character width is changed to ' A_2 ' and the same process is repeated and vice-versa. Even if the match is not encountered, the average character width is further changed to ' A_3 ' and vice versa. This process is continued till the correct segmentation boundary between the two characters is identified. Here A_1, A_2 and A_3 are the predetermined average character widths.

Algorithm: Recursive Segmentation and Recognition

1. *Input an array L_i of segmented lines*
2. *Initialize average character width array $A(k)$ with A_1, A_2 and A_3*
3. *Initialize $k=1$*
4. *For each $L_i \in$ document D*
 - a. *Compute the vertical histogram H .*
 - b. *Segment the line L_i into words W_j*
 - c. *For each W_j*
 1. *Traverse to right most column 'n'.*
 2. *Extract the segment ranging from n th column to $n-A(k)$.*
 3. *Compare the moment features of extracted block with character classes C_s*
 4. *If $n-A(k) \notin C_s$ then $K=k+1$*
Go to step 2
- else*
 1. *Mark segmentation boundary as $n-A(k)$ th column*
- Done*
- Done*
- Done*
5. *Stop*

The figure 3 and figure 4 depicts the results of stage three of the algorithm.

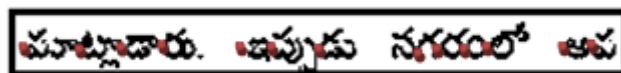


Figure 3: The Segmented Lines With Boundaries Marked For Segmentation



Figure 4: Segmented Characters As Per The Boundaries Located

C. Central Moments:

An image moment is a certain particular weighted average (moment) of the image pixels' intensities, or a function of such moments, usually chosen to have some attractive property or interpretation. An image moments include area (or total intensity), its centroid, and information about its orientation.

Central moments are defined.

$$\mu_{pq} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \bar{x})^p (y - \bar{y})^q f(x, y) dx dy$$

where $\bar{x} = \frac{M_{10}}{M_{00}}$ and $\bar{y} = \frac{M_{01}}{M_{00}}$ are the components of the centroid.

The proposed methodology of recursive approach of segmentation and recognition has produced quite promising results as evident from figure 3 and figure 4. The algorithm is solely dependent upon the average character widths computed and the database of central moments of classified character groups [16, 17]. The increase in number of character widths may further improve the efficiency in segmentation process and also lets the approach to handle varying font sizes and styles.

Experimental Analysis

The algorithm has been experimented with more than 150 images which include printed documents from text books, various newspapers and other documents. The accuracy of the algorithm depends solely on the frequency of changes of the average character widths during the identification of segmentation boundaries between the touching characters. The best performance of the algorithm is attained only when the segmentation boundaries are identified with the first average width chosen every time. The accuracy of the algorithm is defined as number of correctly segmented characters

to the total number of characters available in the document image. The algorithm for segmentation of overlapping lines gives extremely good results only when there are no touching consonants from previous line to the next line. The algorithm can be further improved in this regard in order to tackle the touching lines. The recursive approach of segmentation and recognition can be further extended to work on various font sizes, since the average character widths are computed only with respect to a standard font sizes used in newspapers and printed text books, which could be further improved by redefining the average character widths. The table 1 depicts the experimental results of various types of documents used to segment the touching characters.

Table 1: Results Obtained

| Document Type | Number of documents | Number of lines | Accuracy |
|----------------------|----------------------------|------------------------|-----------------|
| Newspaper | 25 | 150 | 93.20% |
| Printed text book | 30 | 130 | 97.07% |

Conclusion

The experimentation conducted with documents of type newspapers and printed text books had obtained an overall accuracy of around 93-97%. The algorithm can produce best results for the two types of documents considered. However when experimented with other types of printed Telugu documents we are able to get 88% of the accuracy. The average elapsed time required to process a document containing 6 lines of 10-15 words on an average is around 0.04445 seconds. The proposed methodology can be further enhanced by applying connected component analysis on the segmented and recognized components in order to separate the subscripts from the compound characters. The good performance of any character recognition experimentations always comprise with the computational complexities due to the wider processing of text.

References

- [1] Abdul Rahiman M.A, Rajashree M.S., "A Detailed Study and Analysis of OCR Research in South Indian Scripts", International Conference on Advances in Recent Technologies in Communication and Computing, 2009. ARTCom '09, 978-0-7695-3845-7.
- [2] Rinki Singh, Manideep kaur, "OCR for Telugu Script Using Back-Propagation Based Classifier", International Journal of Information Technology and Knowledge Management, July-December 2010, Volume 2, No. 2, pp. 639-643.
- [3] Mamatha H.R, Karthik S, Srikanta Murthy K, "Classifier Fusion Method to Recognize Handwritten Kannada Numerals", PES institute of Technology, Bangalore.

- [4] Mamatha Hosalli Ramappa, Srikantamurthy Krishnamurthy, “Skew Detection, Correction and Segmentation of Handwritten Kannada Document”, *International Journal of Advanced Science and Technology*, Vol. 48, November, 2012.
- [5] J. Bharathi, P. Chandrasekar Reddy, “Segmentation of Touching Conjoint Consonants in Telugu using Minimum Area Bounding Boxes”, *International Journal of Soft Computing and Engineering (IJSCE)*, ISSN: 2231-2307, Volume-3, Issue-3, July 2013.
- [6] M. Swamy das, CRK Reddy, A. Govardhan, Sai Krishna, “Segmentation of overlapping text lines, characters in printed Telugu text document images”, *International Journal of Engineering Science and Technology*, Vol. 2(11), 2010, 6606-6610.
- [7] Srinivasa Rao A V, Mary Junitha M, Shankara Bhaskara Rao G, Subba Rao A V, “Segmentation of Touching Telugu Characters under Noisy Environment”, *Journal of Emerging Trends in Computing and Information Sciences*, Vol. 5, No. 9 September 2014 ISSN 2079-8407.
- [8] Mamatha H R, Srikanta Murthy K, “Morphological Operations and Projection Profiles based Segmentation of Handwritten Kannada Document”, *International Journal of Applied Information Systems (IJ AIS)* – ISSN : 2249-0868, Foundation of Computer Science FCS, New York, USA, Volume 4– No.5, October 2012.
- [9] Srinivasa Rao A V, D R Sandeep, V B Sandeep, S Dhanam Jaya, “Segmentation of Touching Hand written Telugu Characters by using Drop Fall Algorithm”, *International Journal of Computers & Technology*, ISSN: 2277-3061, Volume 3 No. 2, OCT, 2012.
- [10] Nallapareddy Priyanka, Srikanta Pal, Ranju Mandal, “Line and Word Segmentation Approach for Printed documents”, *IJCA Special Issue on “Recent Trends in Image Processing and Pattern Recognition”*, RTIPPR, 2010.
- [11] Monish Rakesh Ajith, “OCR for Telugu script”, *LIB Linear*, December 16, 2011.

