



Segmentation of Offline Malayalam Handwritten Character Recognition

Sangeetha Sasidharan*, Anjitha Mary Paul
Department of Computer Science & MG University
India

Abstract— *Offline handwriting recognition is the process of finding letters and words present in digital image of handwritten text. Because of large variation of writing styles of individuals at different times and among different individuals Offline Handwritten Recognition (OHR) turn to be most interesting and challenging task. This paper focuses on to segmentation part in offline Malayalam handwritten character recognition.*

Keywords— *OHR, Binarization, Median filtering, Horizontal projection profiles.*

I. INTRODUCTION

Offline handwriting recognition is the process of finding letters and words present in digital image of handwritten text. Due to various applications such as postal automation, form processing, storing large volume of manuscripts into digital form and reading aid for blind, offline handwritten recognition (OHR) become a more popular research area. Because of large variation of writing styles of individuals at different times and among different individuals OHR turn to be most interesting and challenging task. The research and development is well progressed for the recognition of the machine-printed documents. In recent Years, the focus of attention is shifted towards the recognition of handwritten script too. Here it primarily focuses on to segmentation part in offline Malayalam handwritten character recognition. ie, Segmentation of untouched characters. The major categories of character recognitions are shown in fig.1.

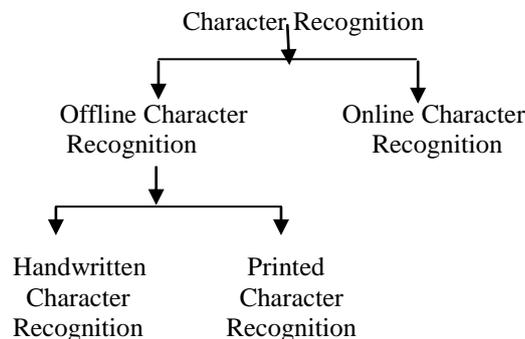


Fig 1: Different classification of character recognition

A. Offline Handwriting Recognition

Offline handwriting recognition involves the automatic conversion of text in an image into letter codes which are usable within computer and text-processing applications. The data obtained by this form is regarded as a static representation of handwriting. Offline handwriting recognition is comparatively difficult, as different people have different hand writing styles. And, as of today, OCR engines are primarily focused on machine printed text and ICR for hand printed text. There is no OCR/ICR engine that supports handwriting recognition as of today.

B. Malayalam Character Set

Malayalam is one of the four major Dravidian languages of South India and one among the twenty two scheduled languages of India with official language status in the State of Kerala and Union territories of Lakshadweep and Mahe, spoken by around 3.5 crores of people and ranked eighth in terms of the number of speakers. Malayalam script is derived from the Grantha script, an inheritor of olden Brahmi script. It is in close propinquity to Tamil and has indelible impression of Sanskrit. It also has the influence of Arabic. Consequently, Malayalam language is enriched with largest number of characters among all Indian languages. And many characters are distinct just with a small variation in appearance. It is syllabic in nature and alphabets are classified into vowels and consonants. Conjunct symbols are used to combine certain consonants. Malayalam language script consists of 15 vowels (Fig 2) and 36 consonants (Fig 3). Although even though Malayalam script has been standardized, people still used to write in both old script and new script.

A. Noise Removal

Median Filter is used to remove the noise present in the image. The median filter is a non-linear digital filtering technique, often used to remove noise from images or other types of signals. The basic idea of this is to examine a sample of the input and decide if it is representative of the signal. This is performed using a window consisting of an odd number of samples. The values in the window are sorted into numerical order. The median value, the sample in the centre of the window, is selected as the output. The oldest sample is then discarded and a new sample is acquired, and the calculation is repeated. The noise removed image is shown below.

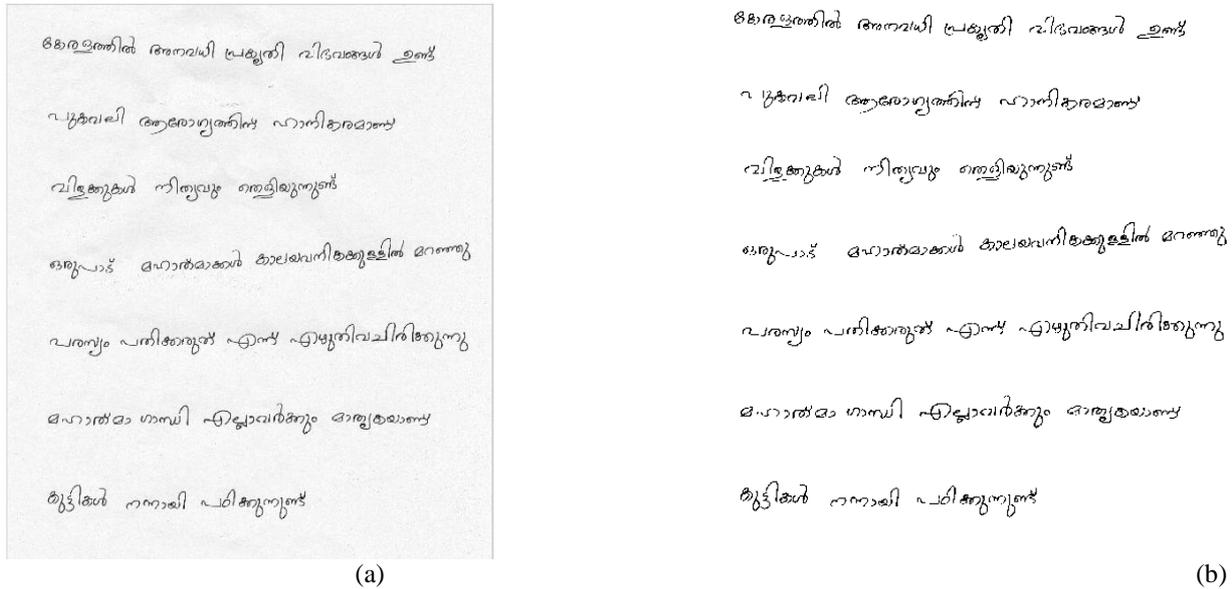


Fig.5 (a) input image (b) noise removed image

A. Binarization

Binarization is a process where each pixel in an image is converted into one bit and you assign the value as '1' or '0' depending upon the mean value of all the pixel. If greater than mean value then its '1' otherwise its '0'. Otsu' Technique is used for binarization which converts gray scale image into binary image. Thus the objective of binarization is to mark pixels that belong to true foreground regions with a single intensity and background regions with different intensities. In Otsu' Technique, It stores the intensities of the pixel in an array. The threshold is calculated by using total mean and variance. Based on this threshold value each pixel is set to either 0 or 1. i.e. background or foreground. Thus here the change of image takes place only once. The binarized image is shown below.

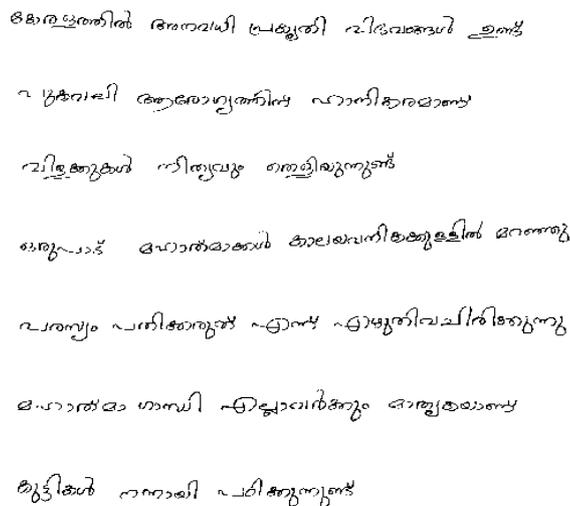


Fig 6: Binarized image

B. Line Segmentation

To segment the handwritten lines, from the document image, the horizontal projection profile is calculated. The horizontal projection profile is the histogram of the number of intensity values of the pixels along every row of the image.

The space between handwritten lines is used to segment the lines. The projection profile will have histogram of zero height between the text lines. Line segmentation is done at these points. The algorithm and the output is shown in figure.

Algorithm 1: Line Segmentation using Horizontal Projection Profile

1. Construct the Horizontal Histogram for the image
2. Find out sum of the white pixel in each row.
3. Using the Histogram, find the rows containing no white pixel.
4. Replace all such rows by 1
5. Invert the image to make empty rows as 0 and text lines will have original pixels.
6. Mark the Bounding Box for text lines
7. Copy the pixels in Bounding Box and save in separate file.

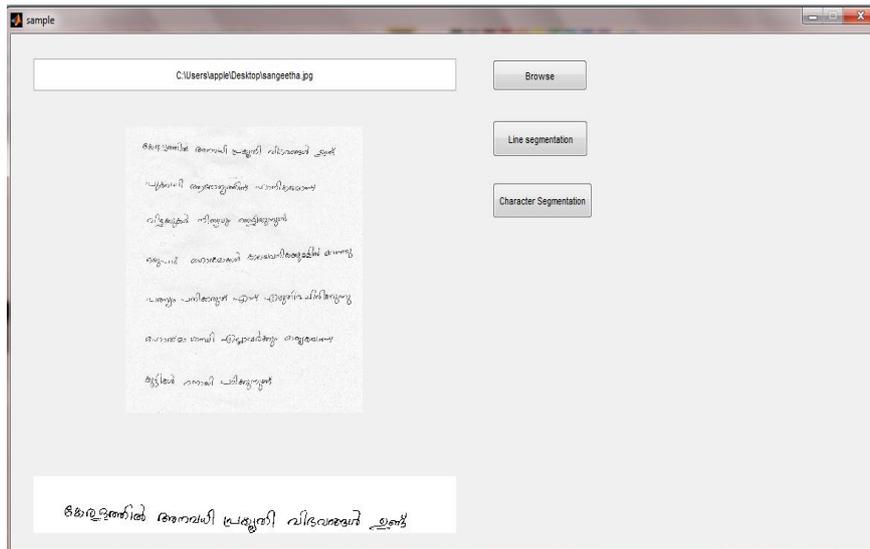


Fig 7: Line segmented image

C. Character Segmentation

In segmentation primarily focus on to Segmentation of untouched characters, Segmentation of consonants touching to valli and Segmentation of consonants touching to chandrakala

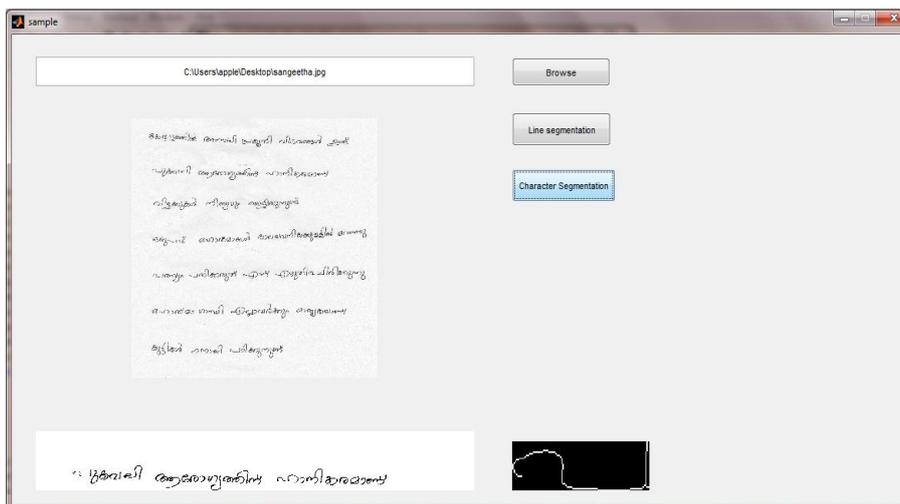


Fig 8: Segmentation of untouched characters

Algorithm2: Segmentation of consonants touching to valli

- Step 1:** Initially, endpoints of the components are calculated
- Step 2:** Find out the junction point by traversing from the last end point
- Step 3:** Check whether the junction point is greater than three by fourth of the height of the component if yes go to step 4 else go to step 5
- Step 4:** Set pixel value of junction point as zero
- Step 5:** Stop

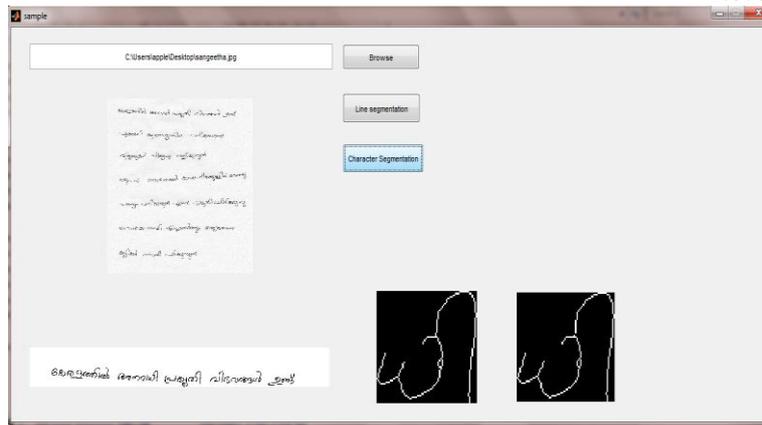


Fig 9: Segmentation of consonants touching to valli

Algorithm3: Segmentation of consonants touching to chandrakala

- Step1:** Initially, endpoints of the components are calculated
- Step 2:** Find out the endpoints that is greater than three by fourth of the height of the component
- Step 3:** Find out the junction point by traversing from those two endpoints
- Step 4:** Set the pixel value of junction points as zero
- Step 5:** Using $y=mx+c$ plot the slope based on the above two points
- Step 6:** Stop

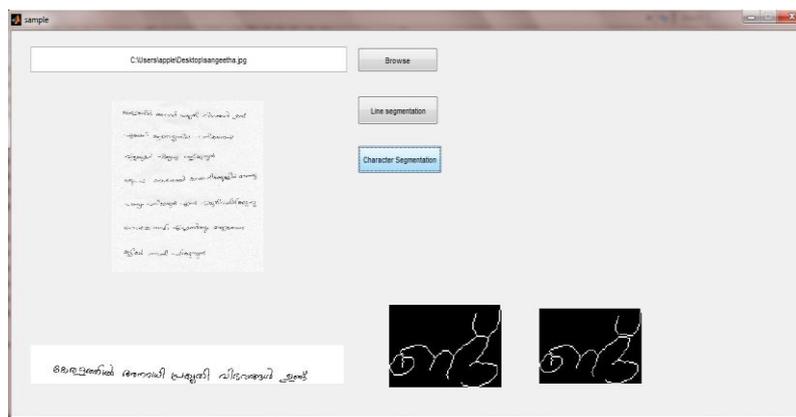


Fig10: Segmentation of consonants touching to chandrakala

VII. RESULTS AND DISCUSSION

Line and character segmentation based on horizontal projection profile and connected component labelling. We have given scan document image as input. After pre processing the input binarized image is segmented into lines. Each line segmented image is again segmented into characters. We could successfully segment 827 out of 879 characters and my proposed work has acquired an efficiency of 94.084187%.

VIII. CONCLUSION

Segmentation of line and characters has been implemented. Here using horizontal projection profile for line segmentation and connected component labelling has been implemented for character segmentation. We have successfully segment consonants touching to valli and chandrakala. It is very easy to segment untouched characters using this method. The main challenge of this method is to segment the touching characters. The future studies are directed towards segmentation of touching characters in a better way.

REFERENCES

- [1] Nobuyuki Otsu (1979), "A Threshold Selection Method from Gray-Level Histogram IEEE Transaction on Systems", Man and Cybernetics.
- [2] Nucharee Premchaiswadi,"A Scheme for Salt and Pepper noise Reduction and Its Application for OCR System", WSEAS Transactions on Computers.
- [3] R Rajeev Kunte and Sudhakar, "A Two Stage Character Segmentation Technique for printed Kannada Text GVIP" Special issue on image sampling and segmentation, 2006.
- [4] A. Zahour, B. Taconet, P. Mercy, and S. Ramdane, "Arabic Hand-written Text-line Extraction", in Proceedings of the Sixth International. Conference on Document Analysis and Recognition, ICDAR 2001, Seattle, USA, September 10-13 2001, pp. 281_285.

- [5] Vignesh J Dongre and Vijay H Mankar, “*Devanagiri document segmentation using histogram approach*”, International Journal of Computer Science, Engineering and Information Technology, 2011
- [6] B. Yanikoglu and P. A. Sandon, ‘*Segmentation of Offline Cursive Handwriting using Linear Programming*,’ Pattern Recognition, Vol. 31, No. 12, 1998, pp. 1825-1833.
- [7] Zhang Ping et al. (2000), “*Text document filters using morphological and geometrical features of characters*”, 5th international conference on Signal Processing Proceedings.
- [8] Zaidi Razak, et al (2008), “*Offline Handwriting Text Line Segmentation: A Review*”, International Journal of Computer Science and Network Security.