International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 10 | Oct 2019 www.irjet.net p-ISSN: 2395-0072

Malayalam Text Detection from Natural-scene Images

Shahana I L[1], Thaha H S[2]

[1]Trainer, SunTec Business Solutions Pvt. Ltd., Trivandrum, Kerala, India
[2]Asst. Prof. Dept. of EEE Musaliar College of Engineering Trivandrum, Kerala, India

---

Abstract - *The research on character recognition for Malayalam script dates back to 1990's. Compared to other Indian languages the research and developments on OCR reported for Malayalam script is very less. The input to the system would be the scanned image of a page of text and the output is a machine editable file. The accuracy of OCR depends on the feature extraction method used. This paper describes an effective OCR for natural scene images. For natural scene images text detection should performed before character recognition. Here text detection is performed by using SWT and centroid analysis)*

Key Words: OCR, SWT, Centroid analysis, HLH patterns, Curvature index

1. INTRODUCTION

Optical Character Recognition (OCR) is the process of conversion of scanned images of handwritten or machine printed documents into computer processable codes. Normal scanning techniques simply create an image of the document, while OCR actually recognizes the print and stores it in an editable text-based format. Current Optical Character Recognition (OCR) techniques can only handle text against a plain monochrome background and cannot extract text from a complex or textured background. In such images Text detection should performed prior to the character recognition.

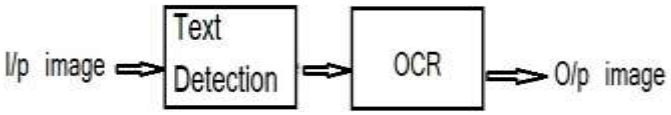

Figure 1: OCR for natural scene images

Since the text data can be embedded in an image or video in different font styles, sizes, orientations, colors, and against a complex background, the problem of extracting the candidate text region becomes a challenging one. Text detection for Malayalam written images is more complex due to varying length of Malayalam characters. So length based comparison is impossible.

Many approaches have been proposed for detecting text from natural scene images, But none of them provide good accuracy for Malayalam. Spatial cohesion refers to the fact that text characters of the same string appear close to each other and are of similar height, orientation and spacing. Two of the main methods commonly used to determine spatial cohesion are based on edge and connected component features of text characters. [1,2] describes the edge-based text extraction and [3] describes Connected Component based text region extraction.[4] describes text detection using SWT(Stroke width transform). Stroke is defined as a uniform region with bounded width and significant extent.

The accuracy of OCR is mainly depends on the feature extraction method used. Some Feature extraction method used in online character recognition are Vertical & Horizontal Line Positional Analyser Algorithm[6], Chain codes Histogram[7] and HLH intensity patterns [8] [5] describe an effective Malayalam OCR with loop, line and curve features.

2. Proposed System

To recognize characters from natural scene images, text detection should be applied prior to the character recognition. We cannot use conventional length and size comparison due to varying length of Malayalam characters. So the accuracy of text detection will be less if we are using connected component or edge based method.

Following flow chart shows text detection of Malayalam characters.

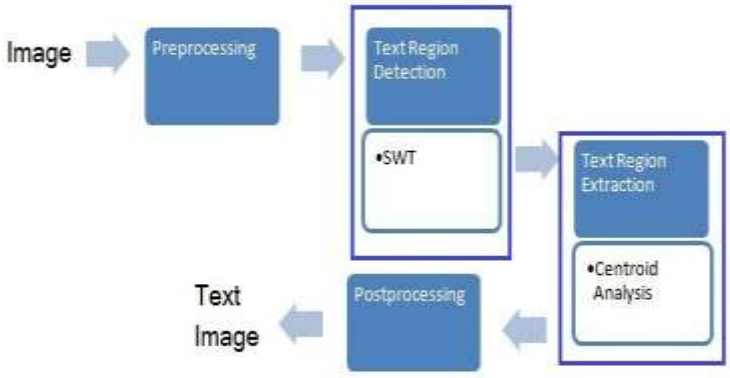

Figure 2: Steps of text detection

Preprocessing

In this step remove the connected components which have aspect ratio either too small or large. Then compute height ratio of various connected component. For a connected component A, If ratio is too small or large for more than half of the other components then remove A.

Text Region detection

Text region detection is done by using stroke width transform (SWT). The stroke width transform (SWT) is a local image operator which computes per pixel width of the most likely stroke containing the pixel. The output of the SWT is an image of size equal to size of the input image where each element contains the width of the stroke associated with the pixel [1].

SWT is computed as follows: The initial value of each element of the SWT is set to infinity. In order to recover strokes, we first compute the edge in the image using Canny edge detector. After that, gradient direction dp of each pixel p considered. If lies on a stroke boundary, then dp must be roughly perpendicular to the orientation of the stroke. We follow the ray r= p + n*d_{p}*image_type, n>0 until another edge pixel q is found. Image type=-1 if it is dark back grounded else +1. We consider then the gradient direction d at pixel q. If dp is roughly opposite to dq each element of the SWT output image corresponding to the pixels along the segment [p q] is assigned width | p-q| unless it already have lower value. Otherwise, if the matching pixel q is not found or if d is not opposite to dp, the ray is discarded.

In text region detection following steps are followed:

1. A=SWT(image)
2. Find largest connected component in A
3. B=SWT in negative direction
4. Reg=A+B
5. Fyn_reg=Reg-small connected component

Here the small connected component will be the connected component which is having size less than largest connected component in A.

Text Extraction

Text extraction is done by using centroid analysis of each connected components in the output of text region detection. Initially, separate image is formed for each connected component. Then for each image replace white pixels with pixels in the output of SWT and centroid analysis is performed.

In centroid analysis[18], the centroid of a connected component should be inside the horizontal range of the other connected component. If Y0 is the y-coordinate of the centroid of one connected component, and y1,y2 are the upper side and bottom side of the bounding box of the other connected component, they should satisfy Y0>= y1+(y2-y1)/3 and Y0<= y2-(y2-y1)/3. It ensures that the two connected components stay in horizontal alignment Second, the height ratio of the two connected components should be greater than 0.83 and less than 1.2. It ensures that they have similar sizes.

Then count the number of character (connected component) which satisfy the centroid analysis. If it is less than the half of the total number of connected components then discards that connected component.

Post processing

In this step morphological erosion and dilation is applied . Then replace white pixels with pixels in the original image. Finally binarise the image. Post processing will helps to remove small blobs or noise and hence give a more accurate and clean image. A problem with this method is that obviously not just text occurs in clusters next to each other and hence anything clustered close to each other will appear in the final map.

Figure:3 shows an example.(a) is input image. (b) is output of edge detection.(c) shows the result of preprocessing. (d), (e), (f) are various stages algorithm. (d) is the image after applying SWT and (e) is the image after applying SWT in negative direction.(f) shows the sum of SWT in negative and positive direction. (g) is the output of text region detection and (h) is output of text region extraction and (i) is the image after post processing

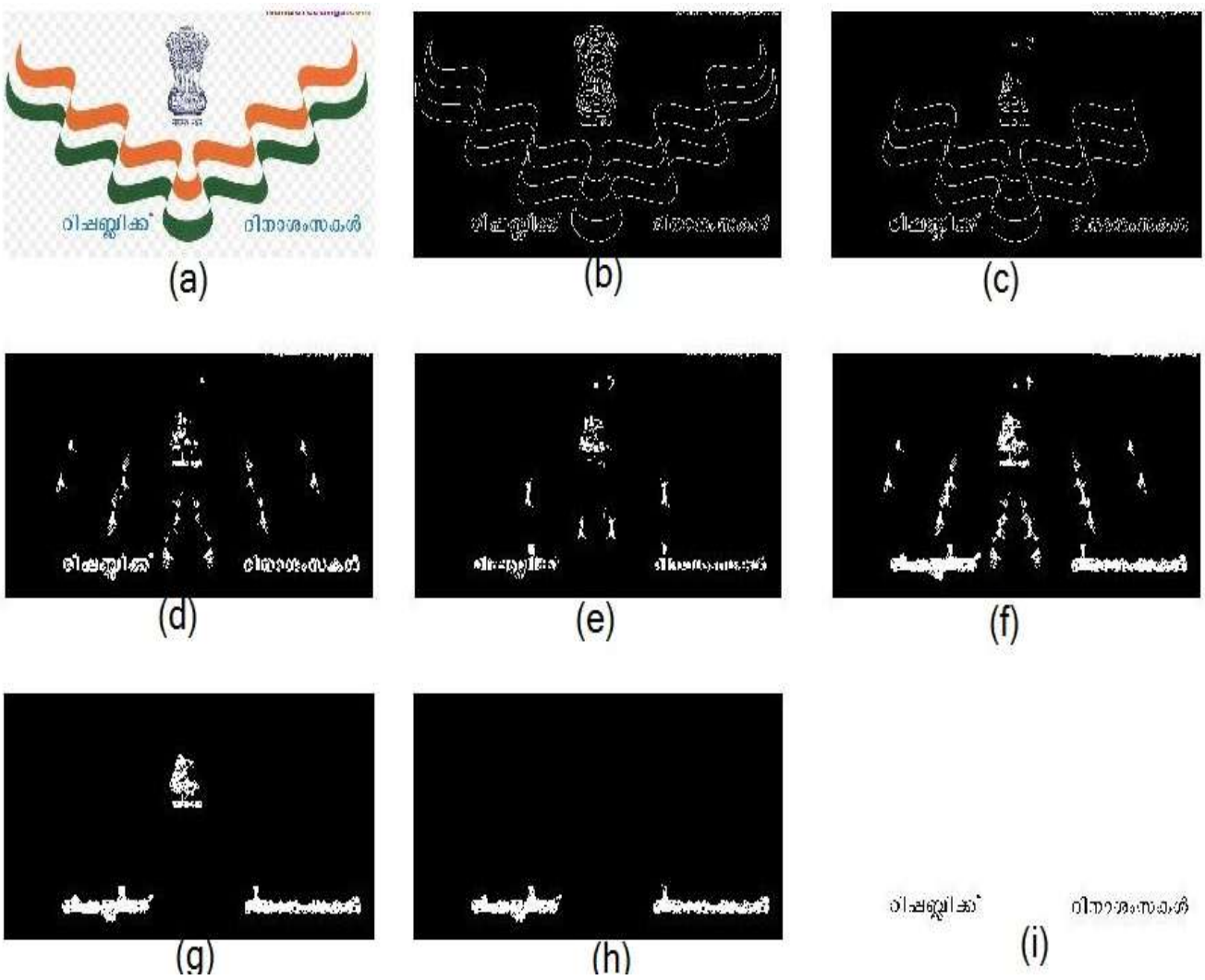

Figure 3: Sample stages of text detection

Character recognition

Character recognition of Malayalam characters can be effectively done using line, loop and curve features. Accuracy of existing OCR can be improved by modifying curve features. Curvature index will represent the curvature associated with each point on the boundary in an absolute way, without distinction of its relative concavity or convexity. This can be included by computing

$\rho 2 = |\rho inv - 1| * Sign$

ρinv is curvature index and Sign is +1 for convex angles and -1 for concave angles.

3. Results

Text detection is performed for different type of images and precision and recall rate is computed.

Type	Image no.	Precision Rate%	Recall rate%
Dark_on_light	1	100	100
	2	100	91.8
	3	91.4	83.1
Light_on_dark	4	100	100
	5	100	93.7
	6	95.4	94.6
Average		97.8	93.8

Thus average precision rate obtained is 97.8% and recall rate is 93.8%. The text detection will not work well for the following type of images:

- An image which is having text with dark background and light background (Stroke width transform will give incorrect results).
- An image has not just text occurs in clusters next to each other (hence anything clustered close to each other will appear in the final image)
- The text in the image is written in circular or highly skewed (Centroid analysis will not detect such type of text)

4. CONCLUSION

Recognition of Malayalam has 2 stages: Text detection, Character recognition. Text detection of Malayalam is difficult compared to other languages. Text detection using dual SWT will give good accuracy. The accuracy of OCR can be improved by including detection of type of angle together with curvature index

REFERENCES

[1] Xiaoqing Liu and Jagath Samarabandu, An Edge-based text region extraction algorithm for Indoor mobile robot navigation, Proceedings of the IEEE, July 2005.
[2] Xiaoqing Liu and Jagath Samarabandu, Multiscale edge-based Text extraction from Complex images, IEEE, 2006.S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, “A novel ultrathin elevated channel low-temperature poly-Si TFT,” *IEEE Electron Device Lett.*, vol. 20, pp. 569–571, Nov. 1999.
[3] Julinda Gllavata, Ralph Ewerth and Bernd Freisleben, A Robust algorithm for Text detection in images, Proceedings of the 3rd international symposium on Image and Signal Processing and Analysis, 2003.
[4] B. Epshtein, E. Ofek, and Y. Wexler,”Detecting text in natural scenes with stroke width transform," IEEE Conference on Computer Vision and Pattern Recognition, 2010
[5] Shahana.i.l,syama.K,Shajeesh.K..U “An effective Malayalam OCR with loop,line and curve features”,ICC2014
[6] M S Rajasree Abdul Rahiman M, "Recognition of handwritten malayalam characters using vertical& horizontal line positional analyzer algorithm",lEEE 2011
[7] Kannan Balakrishnan Jomy John, Pramod K. V, "Offline handwritten malayalam character recog-nition based on chain code histogram",IEEE 2011
[8] Aswathy Shajan M Abdul Rahiman, "Isolated handwritten malayalam character recognition using HLH intensity patterns",Second Inter-national Conference on Machine Learning and Computing 2010