

# Segmentation Of Devanagari Handwritten Text Using Thresholding Approach

Vijay More, Madan Kharat, Shyamrao Gumaste

**Abstract:** Image segmentation is one of the most important and required step in handwritten Devanagari text recognition process. Accuracy of an algorithm, which works on features extracted from segmented characters, highly depends on proper segmentation of handwritten text. There is sufficient segmentation work carried out on European and Chinese text datasets, whereas segmentation on Devanagari text still needs to address in more detail because segmentation and recognition of Devanagari text is difficult due to complex orientation of characters, connected characters and presence of shirorekha. After preprocessing like smoothing and noise removal, image dilation is to be performed and then finding of image contours for selecting region of interest as segmentation. Experimentation performed on PHDIndic\_11 dataset and acceptable segmentation accuracy found is 85.12%.

**Index Terms:** Devanagari, preprocessing, segmentation, threshold.

## 1. INTRODUCTION

There are various application areas where handwritten Devanagari script identification is helpful which include document examination, security, financial activity, forensics, and used as access control. Script identification depends on patterns generated by a writer. These patterns are different for different persons who write document. Those patterns are extracted as features. Handwritten script is available on the historical documents, letters written by writers, manuscript by the authors, post cards, etc. Every handwritten document need to be preprocessed before it is given for recognition. Image segmentation is one of the most important and required step in handwritten character recognition processes. Image segmentation technique is used for dividing an image into easily recognizable segments. These small segments are basically a single atomic part of the script. Usually, it is a single character or a component of the character.

### 1.1 Issues and Challenges

**1.1.1. Issues:** Connected composite characters are a major issue. Overlapped handwritten lines of text confuse segmentation methods which increases complexity of segmentation. Required character may be ignored due to non connectivity like 'Anuswara' which is a point appearing at the top of a character, 'Ardha Chandra' is a half moon like character, and various alike characters in Devanagari text.

**1.1.2. Challenges:** Increased complexity of segmentation due to large amount of variations in writing style. Overlapped written lines of text may confuse segmentation process from identifying character boundaries correctly. Degraded historical handwritten document segmentation is another great challenge. Limited work in the domain of segmentation on Devanagari text is

another added challenge.

## 2. LITERATURE SURVEY

Babu, Subith et al. [1] worked on end and joint points of an image and maintained graph for recording all the primitive skeleton curves. The vertices of polygon curve are recorded as new nodes of graph G. Two end points of the curve represent connections between two nodes. Same joint points are merged together. All subset nodes of G with 8-neighbors are replaced with a new node which keeps the connections among nodes of those subsets and the other nodes of G and it is assigned the coordinates which is average coordinates of the nodes of the subset. Author found 85% segmentation accuracy over 250 handwritten touching character images. Author claimed that their proposed method do not work for certain characters due to variation in handwriting. Guruprasad, Prathima et al. [2] performed Devanagari script segmentation which produces lines, words, and characters separately using Projection-Profiles based technique. Author used vertical projection profile and horizontal projection profile. Author applied this method both on printed and handwritten documents. Bhattad, Anmol J et al. [3] used fuzzy logic system to segment unconstrained handwritten Devanagari words into characters. Upper part above headline is separated by counting non-black pixel to black and black to non-black. Author segment characters by recording and using location of first non-black pixel for each column, thickness of black stroke appears first, count of non-black pixels, and run length count of non-black pixels. As per the characteristics of the character element one of the functions from triangular, Gaussian, trapezoidal, singleton, or piece wise linear are selected as membership function to detect segmentation. Author found 93.39% accuracy in segmenting Devanagari characters. Thakral, B. et al. [4] claims 95% accuracy in segmenting touching, conjunct characters, and that of 88% for overlapping characters. Author uses Cluster Detection technique by removing header line of the word using horizontal projection profile, then finds gap between characters of the word, and store these values into an array. Then, locate the pixel cluster by getting the mid of all array values for cluster separation. Bhujade, Vaishali G. et al. [5] worked on segmentation of Devanagari handwritten characters in four steps. In first step, author removed header line using horizontal projection method. In second step, upper modifiers are

- Vijay More, Faculty, MET's Institute of Engineering, Bhujbal Knowledge City, Nashik, vbmore2005@rediffmail.com, PH: 9421991569
- Madan Kharat, Head, Dept. of Computer Engineering, MET's Institute of Engineering, Bhujbal Knowledge City, Nashik, mukharat@rediffmail.com
- Shyamrao Gumaste, Head, Dept. of Information Technology, MET's Institute of Engineering, Bhujbal Knowledge City, Nashik, svgumaste@gmail.com