22-23

Volume 66, Issue 2, 2022

**Journal of Scientific Research**

of

The Banaras Hindu University

# Segmentation of Lines and Words of Handwritten Devanagari Text using Connected Components with Statistics Method

Vijay More[1][0000-0002-2037-7134] and Madan Kharat[2][0000-0001-8862-9775]

[1]MET's Institute of Engineering, Bhujbal Knowledge City, Nashik, affiliated to Savitribai Phule Pune University, Pune.
vbmore2005@rediffmail.com

[2]Dept. of Computer Engineering, MET's Institute of Engineering, Bhujbal Knowledge City, Nashik.
mukharat@rediffmail.com

**Abstract. The pre-processing activities for handwritten Devanagari text recognition includes an significant step called Segmentation. The segmentation accuracy of Devanagari text characters depends entirely on the accurately segmented lines and words in the handwritten documents. The process of segmenting lines and words correctly leads to many issues. More detailed information is lagging on the segmentation of lines and words from Devanagari text documents, whereas it is available more for other script documents in the literature. Here, we accomplished the task of segmenting the lines and words using Connected Components with Statistics Method on PHDIndic_11 dataset. Experimentation using above mentioned method resulted in line segmentation accuracy of 91.91% and word segmentation accuracy of 72.89% which outperforms over Global threshold and Otsu's optimum threshold methods.**

**Keywords: Connected component, Devanagari, preprocessing, segmentation, statistics.**

## 1. Introduction

Segmentation is one of the most important and basic tasks of preprocessing during image processing and hence is a basic task in recognition of Devanagari text. The segment is usually a single character or even a part of the character as we are interested to find the pattern of the character.

Handwritten Devanagari text documents are preprocessed first and then recognition is carried out after extracting useful features. Devanagari handwritten document script extraction technique is used in many application domains.

The handwritten Devanagari text segmentation process includes several issues and associated challenges as well. As a result, correct recognition helps in recognizing the correct character.

### 1.1 Issues and Challenges

**Issues.** In Devanagari script, we have connected and composite characters which is a mojor issue as shown in Figure 1. As we are concentrating on handwritten characters, overlapping of lines may maximize the complexity of segmentation as shown in Figure 2. Third and foremost complexity arises due to ignorance caused due to 'Anuswara' which is a point appearing at the top of a character, which may be non-connectivity in nature as in Figure 3. Fourth complexity also arise due to 'Ardha Chandra' as shown in Figure 4 a half-moon-like character, and various alike characters in Devanagari text.

**Challenges.** Apart from the issues discussed above, increased complexity of segmentation due to large amount of variations in writing style shown in Figure 5 is a challenge. Second challenge is due to overlapped hand written lines of text as shown in Figure 2 may confuse segmentation process and leads to issues in 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 which is also a scope for research.

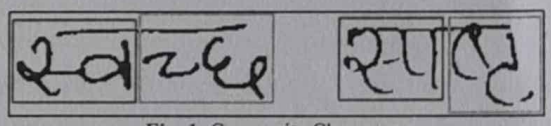


**Fig. 1.** Composite Characters

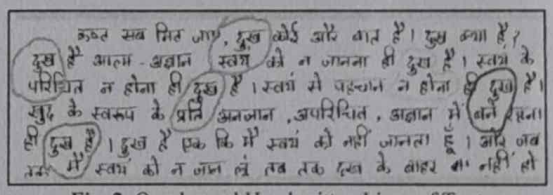


**Fig. 2.** Overlapped Handwritten Lines of Text

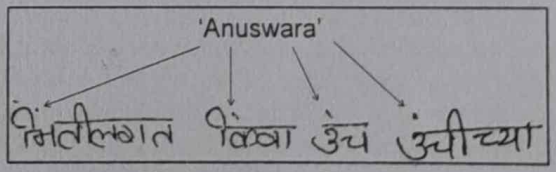
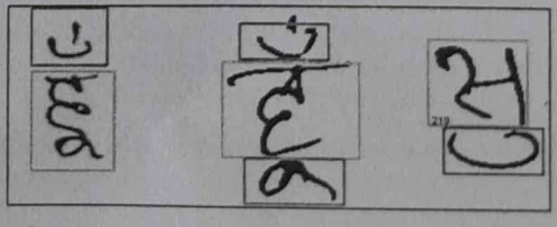



**Fig. 3.** 'Anuswara' Characters





DOI: 10.37398/JSR.2022.660224