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

Vijay More\textsuperscript{1}(0000-0002-2037-7134) and Madan Kharat\textsuperscript{2}(0000-0001-8862-9775)

\textsuperscript{1}MET’s Institute of Engineering, Bhujbal Knowledge City, Nashik, affiliated to Savitribai Phule Pune University, Pune. vbmore2005@rediffmail.com
\textsuperscript{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 a 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 major 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.
2. Literature Survey

Babu, Subith et al. [1] converted handwritten image components of Indic scripts for script identification. These scripts are Bangla, Hindi, Gurmukhi, Bengali, Tamil, Telugu, Malayalam, and Kannada characters. The segmentation accuracy achieved is 93.39% for Devanagari characters segmentation.

Thakral, B. et al. [3] experimented for conjunct and overlapped characters in which the author uses the Cluster Detection technique. The segmentation accuracy is 95% for touching and conjunct characters segmentation and found 88% segmentation accuracy for overlapping characters too.

Bhujade, Vaishali G. et al. [4] experimented with Devanagari handwritten characters segmentation. The segmentation accuracy claimed is 83.02% for half characters in handwritten text.

Kapoor, Shuchi et al. [5] identified joint points and formed bounding boxes covering the identified joint points and then applied segmentation to touching characters based on their height and width. Average segmentation accuracy they found to be 75.75% for Devanagari text.

Palakollu, Saiprakash et al. [6] worked to straighten the header line, and then the upper modifier is separated, along with separation of consonant and lower modifier. The author used the horizontal projection profile technique for segmentation of words and segmentation accuracy claimed is 89.90%.

Garg, Naresh Kumar et al. [7] use structural properties of text to segment the half characters in handwritten Hindi text. An author claiming the achievement of the segmentation accuracy as 83.02% for half characters in handwritten text.

Ladwani, Vandana M. et al. [8] uses morphological operations for segmentation. A neighborhood tracing algorithm was used by the author for segmentation. Segmentation accuracy was found to be 54.66%.

Garg, Naresh Kumar et al. [9] segmentation technique used was based on structure approach for handwritten Hindi text. Vertical projection profile i.e. the histogram of an input image is used, where the zero valley peaks indicate space between words and characters. Accuracy claimed by the author is 79.12% on the segmentation of characters.

Sarkar, Ram, et al. [10] feature-based approach to identify Matra pixels from a word image, and the value so obtained is normalized to the maximum longest run value of any pixel within the word image. The author proposed bell-shaped membership functions to map the horizontalness feature values of each row to determine its belongingness in the Matra region. The segmentation accuracy achieved is 94.8%.

Garg, Naresh Kumar et al. [11] author worked on the method which detects header line, detects baseline, and follows the contour. Skew correction, thinning, or noise removal was not done on the data. The method is applicable only for line segmentation. Line segmentation accuracy found by the author is 97% on handwritten Gurmukhi script. Since character segmentation is not addressed, the paper may help in limited scope to the character segmentation domain.

Otsu Nobuyuki [12] proposed a new terminology in segmentation that he suggested using the optimum threshold selection method using Gray-Level Histograms. He suggested automatic threshold selection as an unsupervised method for image segmentation. An optimal threshold is picked up using the non-parametric method by the discriminant criterion from the gray-level histogram.

Sk Md Obaidullah et al. [13] provided PHDIndic_11: page-level handwritten document image dataset of 11 official Indic scripts for script identification. PHDIndic_11 is a page-level handwritten document image dataset of 11 official Indic scripts for script identification. These scripts are Bangla, Devanagari, Roman, Urdu, Oriya, Gurumukhi, Gujarati, Tamil, Telugu, Malayalam, and Kannada. Devanagari scripts of 220 different documents written by different writers from this data set are considered for the proposed work.

More Vijay et al. [14] studied and reviewed various techniques of handwritten Devanagari text recognition. The author performed a comparative study of various techniques which can be extended further by the researcher community.

Vijay More et al.[15] observed that the segmentation accuracy of Devanagari text characters depends on the segmented lines.
and words from handwritten documents. The author identified many issues and challenges for segmenting lines and words from these handwritten documents. Global threshold and Otsu’s optimum threshold methods were experimented by the author and found 85.12% segmentation accuracy.

3. Proposed Approach

In most research cases, the segmentation accuracy for a single component (line, word, character) was observed very low. Instead, researchers mentioned aggregate performance. This work is experimented separately on line and word segmentation. Their results are mentioned separately for a clear understanding of the results obtained. These results are compared with other researchers’ results. In this work, experimentation is done using connected component analysis as an improved method for segmentation of lines and words specifically and their results obtained are analyzed with the results obtained by the author [15].

3.1 Connected Component Analysis

Connected components scan an image and group a set of pixels into components based on pixel connectivity. All pixels in a connected component having similar pixel intensity values and are connected. Once all groups have been determined, each pixel is labeled with a gray-level or a color (color labeling) according to the component it was assigned to. Pixel connectivity describes a relation between two or more pixels. For two pixels to be connected they have to fulfill certain conditions on the pixel brightness and spatial adjacency.

For two connected pixels, their pixel values must belong to same set of values V.

For a grayscale image, V={22,23,...40}, range of graylevel
For a binary image, V={1}, either 1 or 0.

3.2 Adjacency Criteria

For a pixel p with the coordinates (x,y) the set of adjacency pixels are calculated by using 4-neighbor or 8-neighbor formula:

\[ N_4(p)=\{(x+1,y),(x-1,y),(x,y+1),(x,y-1)\} \]
\[ N_8(p)=N_4(p) \cup \{(x+1,y+1),(x,y+1),(x-1,y+1),(x-1,y-1)\} \]

Fig. 7. Adjacency Criteria a) 4-neighbor, b) 8-neighbor

3.3 Statistics includes

- The leftmost (x) coordinate which is the inclusive start of the bounding box in the horizontal direction.
- The topmost (y) coordinate which is the inclusive start of the bounding box in the vertical direction.
- The horizontal size of the bounding box.
- The vertical size of the bounding box.
- The total area (in pixels) of the connected component.

4. Dataset

Dataset used for implementation of the proposed method is PHDIndic_11 [13]. This dataset consists of 220 images of handwritten documents in Devanagari script. This dataset is used by author [15] in their implementation. These handwritten documents are written by different writers. Example handwritten image document is given in figure 8.

![Fig. 8. Sample Handwritten Image Document](image)

These handwritten image documents are 96 dpi with 8 bit depth. Author [15] categorized these documents in seven categories as partial header lines, congested lines, slope documents, no header lines, sloped with congested lines documents, normal image documents and slope with no header lines documents. Results obtained by author [15] is also categorized according to these different categories. Experimentation is carried out on these documents and obtained results are categorized according to these document categories.

5. Results

Document category wise segmentation analysis

a) Congested lines. There are 27 such documents identified which are categories as congested lines. Sample of congested line document is p_dev_0074.tif and is shown in figure 9.

![Fig. 9. Congested line sample document](image)
Figure 10 is the correct with respect to segmentation accuracy, i.e. correctly segmented lines or words with respect to how many number of line and words segmented, which is found 24.68% for line segmentation and 48.83% for word segmentation.

Figure 11 is the correct with respect to actual accuracy, i.e. correctly segmented lines or words with respect to how many number of line and words actual available, which is found 19.42% for line segmentation and 47.92% for word segmentation.

Figure 12 is the Segmented with respect to actual accuracy, i.e. number of lines and words segmented with respect to how many number of line and words actual available, which is found 72.71% for line segmentation and 67.06% for word segmentation.

b) No-header lines. There are 6 documents categories as no-header lines. Sample of no-header line document is p_dev_0176.tif and is shown in figure 13.

Represention of figure 14 is the correct with respect to segmentation accuracy, i.e. correctly segmented lines or words with respect to how many number of line and words segmented, which is found 41.57% for line segmentation and 57.83% for word segmentation.

Figure 15 is the correct with respect to actual accuracy, i.e. correctly segmented lines or words with respect to how many number of line and words actual available, which is found 39.17% for line segmentation and 51.07% for word segmentation.
Figure 16 is the Segmented with respect to actual accuracy, i.e. number of lines and words segmented with respect to how many number of line and words actual available, which is found 92.96% for line segmentation and 73.41% for word segmentation.

c) Partial header lines. There are 9 such documents identified which are categories as partial header lines. Sample of partial header line document is p_dev_0008.tif and is shown in figure 17.

Figure 18 is the correct with respect to Segmentation Accuracy which is found 89.44% for line segmentation and 79.72% for word segmentation.

d) Slant lines. There are about 37 handwritten documents observed and categorized as slant documents. Sample slant document is p_dev_0096.tif and is shown in figure 21.
Figure 22 is the correct with respect to segmentation accuracy which is found 28.31% for line segmentation and 51.84% for word segmentation.

Figure 23 is the correct with respect to actual accuracy which is found 27.35% for line segmentation and 64.26% for word segmentation.

Figure 24 is the Segmented with respect to actual accuracy which is found 87.47% for line segmentation and 61.69% for word segmentation.

e) Slant with congested lines. There are about 9 handwritten documents observed and categorized as slant with congested lines. Sample Slant with congested line document is p_dev_0042.tif and is shown in figure 25.

Figure 25 is the Slant with congested line sample document [p_dev_0042.tif]

Figure 26 is the correct with respect to segmentation accuracy which is found 35.17% for line segmentation and 57.63% for word segmentation.

Figure 27 is the correct with respect to actual accuracy which is found 31.98% for line segmentation and 68.29% for word segmentation.
Fig. 28. Segmented with respect to actual accuracy

Figure 28 is the Segmented with respect to actual accuracy which is found 88.96% for line segmentation and 78.88% for word segmentation.

f) Slant with no-header lines. There are about 20 handwritten image documents observed which are categorized as slant with no-header lines. Sample slant with no-header lines document is p_dev_0214.tif and is shown in figure 29.

Fig. 29. Slant with no-header line sample document [p_dev_0214.tif]

Fig. 30. Correct with respect to Segmentation Accuracy: Lines , Words

Figure 30 is the correct with respect to segmentation accuracy which is found 38.61% for line segmentation and 64.36% for word segmentation.

g) Normal documents. There are about 117 such handwritten image documents observed as normal documents. Sample normal document is p_dev_0217.tif and is shown in figure 33.

Fig. 31. Correct with respect to Actual Accuracy

Figure 31 is the correct with respect to actual accuracy which is found 38.53% for line segmentation and 83.85% for word segmentation.

Fig. 32. Segmented with respect to actual accuracy

Figure 32 is the Segmented with respect to actual accuracy which is found 98.54% for line segmentation and 65.04% for word segmentation.
Figure 34. Correct with respect to Segmentation Accuracy

Figure 34 is the correct with respect to segmentation accuracy which is found 68.02% for line segmentation and 76.34% for word segmentation.

Figure 35. Correct with respect to Actual Accuracy

Figure 35 is the correct with respect to actual accuracy which is found 66.89% for line segmentation and 82.00% for word segmentation.

Figure 36. Segmented with respect to actual accuracy: Lines, Words

Figure 36 is the Segmented with respect to actual accuracy which is found 96.51% for line segmentation and 78.41% for word segmentation.

h) Overall documents. Collection of all category documents is overall documents of the dataset PHDIndic_11 [13] and which are 220 nos and whose samples are already provided in above results discussion.

Figure 37. Correct with respect to Segmentation Accuracy

Figure 37 is the correct with respect to segmentation accuracy which is found 51.67% for line segmentation and 66.54% for word segmentation.
Figure 38 is the correct with respect to actual accuracy which is found 50.06% for line segmentation and 73.23% for word segmentation.

Figure 39 is the segmented with respect to actual accuracy which is found 91.91% for line segmentation and 72.89% for word segmentation.

Accuracy comparison of lines and words between Global thresholding method, Otsu’s thresholding method, and Connected Component with Statistics method as Improved method is shown in Table (1).

<table>
<thead>
<tr>
<th>Document category</th>
<th>Global method for lines</th>
<th>Otsu method for lines</th>
<th>Improved Method Lines</th>
<th>Global method for words</th>
<th>Otsu method for words</th>
<th>Improved Method Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Congested lines</td>
<td>2.70</td>
<td>3.29</td>
<td>24.68</td>
<td>55.13</td>
<td>61.21</td>
<td>48.83</td>
</tr>
<tr>
<td>No header lines</td>
<td>1.85</td>
<td>2.96</td>
<td>41.57</td>
<td>0.83</td>
<td>5.81</td>
<td>57.83</td>
</tr>
<tr>
<td>Partial header lines</td>
<td>8.57</td>
<td>8.22</td>
<td>89.44</td>
<td>40.10</td>
<td>49.75</td>
<td>79.72</td>
</tr>
<tr>
<td>Slant</td>
<td>3.54</td>
<td>4.41</td>
<td>28.31</td>
<td>48.68</td>
<td>58.35</td>
<td>51.84</td>
</tr>
<tr>
<td>Slant with congested lines</td>
<td>0.34</td>
<td>0.15</td>
<td>35.17</td>
<td>63.49</td>
<td>65.97</td>
<td>57.63</td>
</tr>
<tr>
<td>Slant with no header lines</td>
<td>0.21</td>
<td>0.21</td>
<td>38.61</td>
<td>4.29</td>
<td>18.86</td>
<td>64.36</td>
</tr>
<tr>
<td>Normal documents</td>
<td>5.56</td>
<td>5.56</td>
<td>68.02</td>
<td>46.18</td>
<td>61.03</td>
<td>76.34</td>
</tr>
<tr>
<td>Overall</td>
<td>4.12</td>
<td>4.36</td>
<td>51.67</td>
<td>43.25</td>
<td>55.26</td>
<td>66.54</td>
</tr>
</tbody>
</table>

Table 1. Correct with respect to Segmented Accuracy

It is seen from (Table 1) that, improved method performed better in all the document categories. If we observe the case of partial header lines documents, it is found 89.44% segmentation accuracy which is much better than 8.57% and 8.22% accuracy obtained using Global and Otsu’s method respectively by author [15]. In word segmentation also the Improved method outperforms more in case of Partial header lines documents in comparison with other category documents.

The table values from (Table 1) are depicted in chart form in figure 40.

Correct with respect to segmented accuracy comparison (figure 40) shows that Connected Component Method performs better than Global and Otsu's method [15] in most of the document categories.
The table values from (Table 2) are mentioned in chart (figure 41) for more aggregate clarity on results.

![Segmented w/ Actual Accuracy](image)

**Fig. 41. Comparison Chart of Connected Component Method vs Global and Otsu's Method Under Segmented with respect to Actual Accuracy**

Segmented with respect to actual comparison (see figure 41) shows that Connected Component Method outperforms too much better than Global and Otsu’s method [15]. In all the document categories, the peak values of connected component method are very much higher than that of Global and Otsu’s method. Line segmentation accuracy obtained is 91.91% which is too much better than that of using methods used by author [15] and word segmentation accuracy obtained is 72.89% using connected component method which also better than author [15] results.

### 6. Conclusion

The experimental results of the connected component analysis method provide the segmentation of lines and words. In this work, we obtained line segmentation accuracy of 91.91% and word segmentation accuracy of 72.89%. Experimentation results show that the method outperforms on Global threshold [3.51% for lines and 4.75% for words] and Otsu’s optimum threshold method [6.95% for lines and 10.78% for words] as per the results obtained in [15].

This work has more scope to extend to many dimensions like the segmentation of composite characters and overlapped lines. Though the work carried out by various researchers in these areas is appreciated, there is still room for improvement.

### 7. References