

# Connected Component based English Character Set Segmentation

Nitigya Sambyal<sup>1</sup>, Pawanesh Abrol<sup>2\*</sup>

<sup>1</sup>Department of Computer Science & IT, University of Jammu, Jammu, Jammu & Kashmir, India-180006

<sup>2</sup>Department of Computer Science & IT, University of Jammu, Jammu, Jammu & Kashmir, India-180006

Email addresses: sambyal.nitigya@gmail.com, pawanesh.abrol@gmail.com

\* Corresponding author

**Abstract**— Character segmentation is a prerequisite phase in document analysis tasks. It helps in extracting characters which aid in indexing and understanding. The research paper proposes a character segmentation technique based on connected component method. The algorithm is tested along a set of jpeg, png and bmp images over English character set and uses a cluster size of five for identifying the separate characters. It has been observed that the proposed connected component based character segmentation approach gives on average 95.44% accuracy for the various test cases under study.

**Keywords**— *Connected component; Character segmentation,; Normalization ;Thresholding*

## I. INTRODUCTION

Character segmentation is an important phase in document analysis and plays a vital role in character recognition system. Character segmentation system takes text line as input and generates separated characters as an output [1]. For character segmentation techniques like binarization, histogram approach [2], connected component analysis etc. can be used. Connected component based method to the detection and recognition of text which have uneven lighting condition, different sizes and shapes. Edge information [3] is used for the Connected Component generation following which difference between adjacent pixels is used to determine the boundaries of potential characters after quantizing an input text image. Local threshold values can then be selected for each text candidate [4]. In this paper, character segmentation using connected component method with cluster size five is proposed. The main advantage of this technique is it is easy to implement, less computationally intensive, efficient and involves no overhead as required in training.

The research paper is organised as follows: Section 2 briefly overviews various connected component based character segmentation algorithms. Section 3 presents the proposed character segmentation algorithm. Section 4 provides the various results and concerned discussions along with accuracy

of the proposed system and section 5 consists of conclusion along with various challenges and future directions.

## II. RELATED WORK

Connected component based approach is a bottom up approach which proceeds by exploring the similarity between adjacent pixels to identify a separated character. Rodolfo P.dos Santos et al. proposed text line segmentation can be done by morphology and histogram projection. In this process, firstly a Y histogram projection is performed which results in the text lines positions. Threshold is applied to divide the lines in different regions. In order to detect the extreme positions of the text in the horizontal direction, an X histogram projection is applied. Another threshold in the Y direction is used to eliminate false words. Finally, in order to optimize the area of the manuscript text line, a text selection is carried out [5].

Bounded box method given by Vikas J Dongre and Vijay H Mankar can be used for segmenting lines, words and characters in a document. It is based on the pixel histogram approach where text images are converted to binarized images by thresholding using otsu's method [6] and then numbers of text lines are detected using horizontal histogram. This is followed by plotting vertical histogram which enables segmenting words in each text line. This is done by determining the columns containing no

white pixels. Finally, each character in the word is segmented after applying thinning operation on the image. A bounding box is marked around each character and saved in a separate file [7].

A segmentation algorithm is proposed by Ohya wherein documents are segmented into regions by using image segmentation method based on adaptive thresholding. Character candidate regions are selected through checking features under assumptions like width, gray level of text, spatial frequency etc. Thereafter, recognition process is applied to cluster the separate parts of one text character together and extract character pattern candidates [8].

### III. PROPOSED APPROACH

The proposed character segmentation approach is based on connected component method. Fig 1. Shows the implementation of the proposed In this process, the input is first normalized. Normalization involves conversion of input to grayscale followed by thresholding which is performed by determining global threshold. For determining the connected components (or segmented characters) a cluster size  $s$  is selected by careful experimental evaluation. It is observed that at cluster size five stray marks (or noise) are filtered out. Thus, components in resultant mask with cluster size greater or equal to size  $s$  is identified as an individual character. A label matrix is used to determine such connected components. Finally, these segmented characters are displayed as output. The algorithm for the proposed approach is explained as follows:

**Step 1** Compute the global threshold level using Otsu's method. This threshold is used to convert an intensity image to binary image and minimizing the interclass variance of the black and white pixels.

**Step 2** Determine the size of each component in the resultant image mask and remove from the binary image all connected components that have less than specified pixels  $s$ .

**Step 3** Make a label matrix that contains labels for the 8 connected objects found in the above image. Also determine the number of components in it.

**Step 4** Measure properties of each connected object to finally return the smallest rectangle containing the connected component.



FIGURE1. Implementation of proposed connected component approach

### IV. RESULTS AND DISCUSSIONS

The proposed text detection and character segmentation algorithm based on connected component technique is checked for a set of images either taken directly from the web or made manually using various application tools.

TABLE 1. Result of character segmentation algorithm on English character set

Test case number	Input	Output of character segmentation
T1	A BOOK	A B O O K
T2	Script	S c r i p t .
T3	STAR	S T A R
T4	Square	S q u a r e
T5	BIRD	B I R D






The outputs of various test cases with English character set T1 to T5 is shown in Table 1. From Table 1, it is evident that the proposed connected component technique gives good results for various test cases under study. However, in test cases T2

and T3 more than one character are identified as single component. This happens as the adjacent characters are either very closely placed or are overlapping.

### V. ACCURACY

The accuracy of the character segmentation approach is shown in Table 2.

TABLE 2. Accuracy of proposed character segmentation algorithm

Input	Character segmentation		
	Characters in document	Identified characters	Accuracy
	12	13	92.307%
	32	34	94.11%
	15	16	93.75%
	33	34	97.05%
	17	17	100%

### VI. CONCLUSION

In this research paper, character segmentation using connected component method with cluster size five has The formula used to compute the accuracy of proposed connected component based character segmentation technique is given as follows:

$$\text{Accuracy of character segmentation} = (\text{characters in document} / \text{Identified characters}) * 100$$

Where the characters in the document are counted manually and recognised characters are obtained as output of the proposed algorithm. From the table it is observed that the algorithm gives on an average 95.44% for the various test cases under study.

been proposed. The algorithm is tested on a set of jpeg, png and bmp images with printed English character set. Since the character segmentation algorithm operates by exploring the connected components it may not provide accurate results in cases where the adjacent characters are either closely placed or are overlapping. Sometimes dot (.) are identified as noise and discarded. The proposed algorithm gives on an average an accuracy of 94.55% for test cases with printed English character set. In future the proposed algorithm can be extended to character sets like Hindi, Urdu, Arabic etc. which have different writing style and formation.

### REFERENCES

- [1] Nitigya Sambyal, Pawanesh Abrol, "Automatic text extraction and character segmentation using maximally stable extremal regions", in *International Journal of Modern Computer Science* Volume 4, Issue 3, pp. 136-141, 2016.
- [2] Richard G. Casey and Eric Lecollinet, "A survey of methods and strategies in character segmentation", *IEEE Transactions On Pattern Analysis And Machine Intelligence*, Vol. 18, Issue 7, pp 690-706, 1996.
- [3] Li Bin, Mehdi Samiei yeganeh, "Comparison for Image edge detection algorithms", *IOSR Journal of Computer Engineering*, Vol. 2, Issue 6, pp 01-04, 2012.
- [4] Feby Ashraf, and Nurjahan V A, "Connected component clustering based text detection with structure based partition and grouping", *IOSR Journal of Computer Engineering*, Vol.16, Issue 5, Version III, pp 50-56, 2014.
- [5] R. P. Dos Santos, G. S. Clemente, T. I. Ren, and G. D. C. Calvalcanti, "Text line segmentation based on morphology and histogram projection," in *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, 2009, pp. 651-655, 2009.
- [6] Jun Zhang, Jinglu Hu, "Image segmentation based on 2d otsu's method with histogram analysis", *Proceedings of IEEE Computer Science and Software Engineering International Conference*, Vol.6, pp. 105-108, 2008.
- [7] V. J. Dongre and V. H. Mankar, "Devnagari document segmentation using histogram approach," *International Journal of Computer Science Engineering And Information Technology*, Vol. 1, Issue 3, pp. 46-53, 2011.
- [8] K. Venkateswarlu and S. M. Velaga, "Text detection on scene images using MSER," *International Journal of Research in Computer And Communication Technology*, Vol. 4, Issue 7, pp. 452-456, 2015.

