International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 4 Issue: 9 50 - 51

Segmentation of the overlapping Kannada Characters

Soumyadeep Sinha
Department of Information Science and Engineering
Dayananda Sagar College of Engineering
Bengaluru-560078, India.
soumyadeep.sinha2011@gmail.com

Abstract— Kannada is a widely spoken language in the southern part of India. Character segmentation of Kannada text is difficult, since adjacent characters in Kannada sometimes overlap in the vertical projection profile. In such cases, the usual method of character segmentation using projection profile is not efficient. In this paper we present a segmentation method in which overlapped characters are separated by connected component analysis.

Keywords-*Overlapped Characters, Vertical Projection, Optical Character Recognition, Connected Component Analysis*

*****

I. INTRODUCTION

Optical Character Recognition (OCR) systems have been effectively developed for the recognition of non-Indian languages. Efforts are on the way for the development of an efficient OCR system for Indian languages, especially for Kannada, a popular South Indian language, which is very rich in alphabets. For developing an effective OCR system, having an algorithm to carefully segment characters into individual components is mandatory to increase the recognition rate.

The classical segmentation approach uses several methods of segmentation such as White Space and Pitch segmentation algorithm and Projection Analysis. The most common approach is to use Projection Profile Analysis since it is simple and fast.
In Kannada text, for words having bottom extension characters (Vatthus), the space between two adjacent characters does not have zero spaced valleys in the vertical projection profile, which makes it difficult to extract individual characters from the word as evident from characters surrounded by the red border in Figure (1)

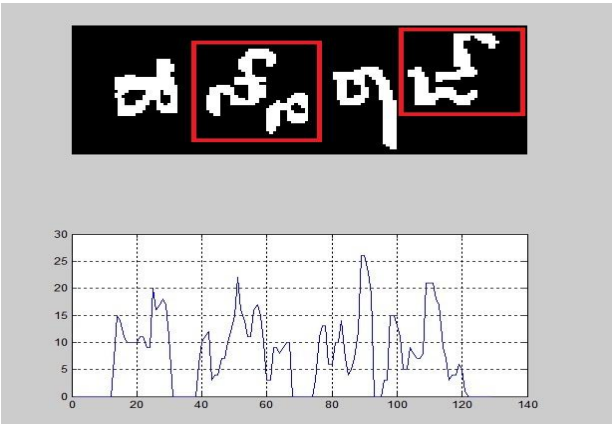

Figure 1

In such situations, we are going for a segmentation technique called the connected component method, which treats each of the individual characters in the text (both main and Vatthu characters) as separate components of the image. In this paper we have described a two-stage method: In the first stage the subsets of connected components are uniquely labeled. The second stage requires segmenting the characters according to their respective labels.

II. SEGMENTATION ALGORITHM

First Stage- Connectivity checks are carried out by checking neighbor pixels' labels (neighbor elements whose labels are not assigned yet are ignored), or say, the North-East, the North, the North-West and the West of the current pixel (assuming 8-connectivity). 4-connectivity uses only North and West neighbors of the current pixel. The following conditions are checked to determine the value of the label to be assigned to the current pixel (4-connectivity is assumed)
Conditions to check:

1) Does the pixel to the left (West) have the same value as the current pixel?
   - **Yes** – We are in the same region. Assign the same label to the current pixel
   - **No** – Check next condition
2) Do both pixels to the North and West of the current pixel have the same value as the current pixel but not the same label?
   - **Yes** – We know that the North and West pixels belong to the same region and must be merged. Assign the current pixel the minimum of the North and West labels, and record their equivalence relationship
   - **No** – Check next condition
3) Does the pixel to the left (West) have a different value and the one to the North the same value as the current pixel?
   - **Yes** – Assign the label of the North pixel to the current pixel
   - **No** – Check next condition
4) Do the pixel's North and West neighbors have different pixel values than current pixel?
   - **Yes** – Create a new label id and assign it to the current pixel

The algorithm continues this way, and creates new region labels whenever necessary as shown in figure 2.

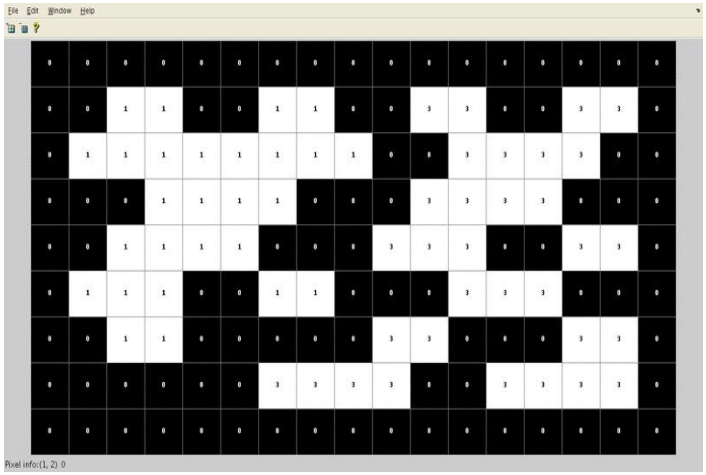

Figure 2

Second Stage- In second stage, the pixels having one label are grouped together and the pixels having another label are grouped differently. Thus the overlapping characters can be differentiated into two parts according to their respective labels as shown in figure 3.

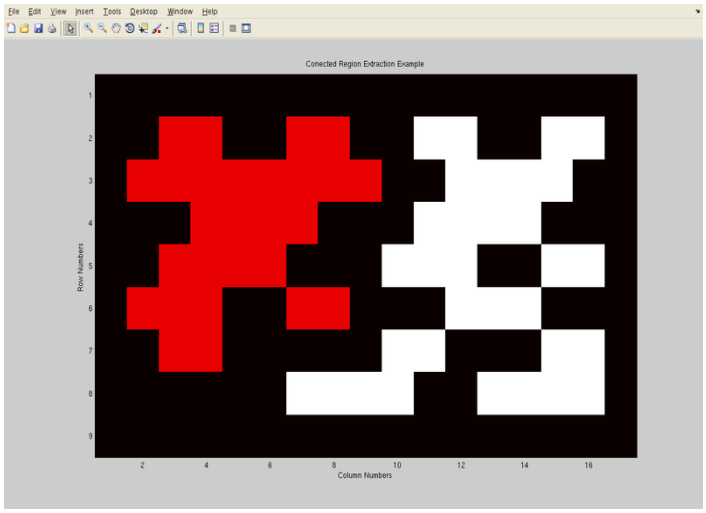

Figure 3

III. RESULTS

The result of the character segmentation of handwritten Kannada text is shown by the following figures. Figure 4 shows the overlapping characters whereas figure 5 shows the segmented characters.

Figure 4

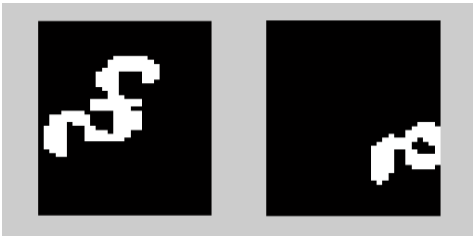

Figure 5

IV. CONCLUSION

In this paper, a new method of generating curved segmentation paths is proposed. A connected component analysis algorithm is presented to segment overlapped characters. From the above figures, it is clear that the segmentation techniques developed for character segmentation produced exceptional results for words written in Kannada. The character separation technique explained above can also be applied to other Indian scripts. However, there are few drawbacks to this technique: The segmentation problem occurs where characters touch each other or in some cases where lower modifier is very small or not forming the loop. Most of the half characters are also segmented correctly but work for properly segmenting the half characters is still in progress.

V. REFERENCES

[1]. S. Mori, C. Y. Suen and K. Yamamoto, "Historical review of OCR Research and development", Proceedings of the IEEE , Vol. 80(7), pp. 1029-1058, 1992.

[2]. K. Sethi, "Machine recognition of Constrained hand printed Devanagari", Pattern Recognition , pp.69-75, 1977.

[3]. A. Zahour, B. Taconet, P. Mercy, and S. Ramdane, "Arabic Hand-written Text-line Extraction", Proceedings of the Sixth International. Conference on Document Analysis and Recognition, ICDAR, Seattle, USA, pp. 281–285, 2001.

[4]. N. Tripathy and U. Pal, "Handwriting Segmentation of unconstrained Oriya Text", in the proceedings of International Workshop on Frontiers in Handwriting Recognition, pp. 306–311,2004.

[5]. T. Yamaguchi, T. Yoshikawa, T. Shinogi, S. Tsuruoka, and M. Teramoto, "A segmentation method for touching Japanese handwritten characters based on connecting condition of lines," Proc. 6th Int. Conf. Document Analysis and Recognition, pp. 837−841, 2001.

[6]. S. Mori, C. Y. Suen and K. Yamamoto, "Historical review of OCR Research and development", Proceedings of the IEEE , Vol. 80(7), pp. 1029-1058, 1992.

[7]. K. Sethi, "Machine recognition of Constrained hand printed Devanagari", Pattern Recognition , pp.69-75, 1977.

[8]. Casey, R. and E. Lecolinet, A Survey of Method and Strategies in Character Segmentation, IEEE Transaction on PAMI, 18(7), pp. 690-706, 1996.

[9]. U. Pal and Sagarika Datta, (2003) "Segmentation of Bangla Unconstrained Handwritten Text ", Proc. 7th Int. 64, doi:10.1109/SCIS.2007.357670.