RESEARCH REVIEW JOURNALS
NCIDCAS-2018
DEC- 2018
Special Issue

ISSN: 2455-3085 (Online)
RESEARCH REVIEW International Journal of Multidisciplinary
www.rrjournals.com [UGC Listed Journal]

Segmentation of Special Character "∴" from degraded Brahmi Script Documents

*[1]Nagane Aniket & [2]Mali Shankar

[1]*Assistant Professor, Department of Computer Science, MIT ACSC,Alandi(D), Pune, Maharashtra (India)*
[2]*Associate Profressor, Department of Computer Science, Dr. Vishwanath Karad MIT World Peace University, Pune (India)*

ARTICLE DETAILS

Article History
Published Online: 22 Dec 2018

Keywords
Brahmi, segmentaion, recognition, binarization, degraded document

***Corresponding Author**
Email: asnagane[at]mitacsc.ac.in

ABSTRACT

Study in this paper presents a method of segmenting special character "∴" from the degraded Brahmi script documents. Proposed method in this paper is based on the distance between three dots which forms the character. Correct/accurate segmentation is necessary In OCR systems. Incorrect segmentation affects the recognition of characters and ultimately the performance of the system.

1. Introduction

Brahmi is the oldest script used in central Asia and in Indian subcontinent. Brahmi script is the origin of all Indian scripts. Brahmi script was commonly used in the period of 400 B.C to 500 A.D. Literature in written in Brahmi script is found mostly as rock inscriptions in early period of brahmi and later found on tamrapatra and bhurjpatra. Literature found as rock cut edicts is preserve and made available to the historians in the form of estampages. Estampage is an impression of inscription on inked paper. Archeological Survey of India, Mysore maintains the digitized copies of estampages. As brahmi is an oldest script, common people don't have understanding of the script and it is difficult for them to read it. By developing an OCR system for Brahmi script reading and understanding the literature available in Brahmi script will become easy.

In OCR systems, processing the document images and performing the segmentation of characters is very important step. If segmentation of characters is not performed correctly, it affects the recognition of characters and ultimately affeacts the performance of system. More the failure in recognition, poor is the performance of the system.[1,2,3,4,5]

The time span in which the Brahmi script was practiced is of nine hundred years which considerably a long period of time. Because of this many variations are observed in the character set of Brahmi script. The period of emperor Ashoka, is the time span where it has been observed that the script is consistent in its use of character set across the Indian subcontinent.

Figure 1(a) depicts 10 vowels and 1(b) depicts 33 consonants from Ashokan Brahmi script.

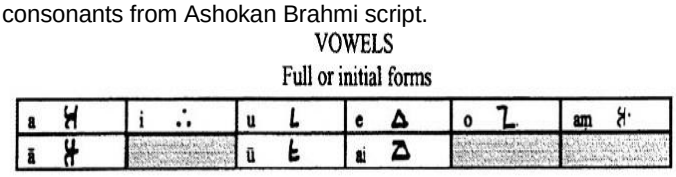

Figure 1(a) Source: Indian Epigraphy by Richard Saloman

From the entire set of alphabetic symbols "∴" is the only character which has three different objects separated from each other. While segmenting the character it is difficult to group the three dots, as they are not connected to each other.

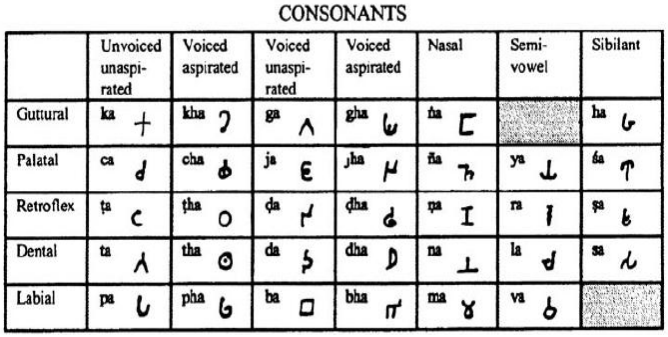

Figure *1(b) Source: Indian Epigraphy by Richard Saloman*

2. Proposed method

Out of entire character set of Ashokan Brahmi script, "∴" is the only character which has collection of three isolated objects as representation of one character symbol. Otherwise all other characters are having symbolic representation with single object as a continuous stroke. While segmenting these characters, it is difficult to segment "∴" character. As its representation contains three isolated objects, the existing methods of segmentation, segments three objects separately and not as group of three objects as desired. Segmenting three dots separately results into different meaning .i.e the single dot symbol is treated as modifier "anuswar". So we expect that the three dots in the characters should be grouped together first and then get segmented as one single character.[6,7,8,9,10] Proposed method for segmenting special character consist of six different steps:

Step 1: Find the average size of characters in image.[11,12]

Find out area of each connected component in an image as:

Avgarea = sum of all objects' area / no of connected components

Where Avgarea is average area of characters in an images

Step II: Find the size of dots.[13]

From manual calculations it is observed that area of dots in an image is strictly less than one third (1/3) of average area

Step III: Store all the dots in an array.

Store all objects into an array having area less than or equal to one third of average area

Step IV: Find the average absolute difference between every three dots with respect to x coordinates and y coordinates.

diffx=(dot1(x)-dot2(x))+(dot1(x)-dot3(x))+(dot2(x)-dot3(x)) (Eqn. 1)

diffy=(dot1(y)-dot2(y))+(dot1(y)-dot3(y))+(dot2(y)-dot3(y)) (Eqn. 2)

avgdiifx = |diffx| / 3 (Eqn. 3)

avgdiffy = |diify| / 3 (Eqn. 4)

where diffx is sum of differences between x coordinates of three dots, diffy is sum of differences between y coordinates of three dots, avgdiffx and avgdiffy is absolute average difference between three dots with respect to x coordinates and y coordiantes.[14,15]

Step V: Find three closed dots/objects

Based on the Distance calculated in previous step, we find three closet dots and group them. From manual calculations it is observed that, if average difference between x and y coordinates is in the range of 12 to 20 then these are three dots belonging to one object "∴"

Step VI: Segment the grouped dots as special character "∴"

To group and segment three closest dots from previous step, we have to find smallest x and y coordinate values from these three dots.Based on these coordinates we should apply new bounding box to segment the character "∴"

3. Experimental Results

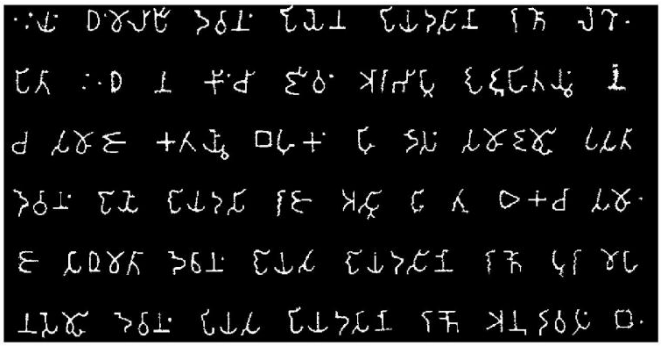

Figure 2: Binarized input image

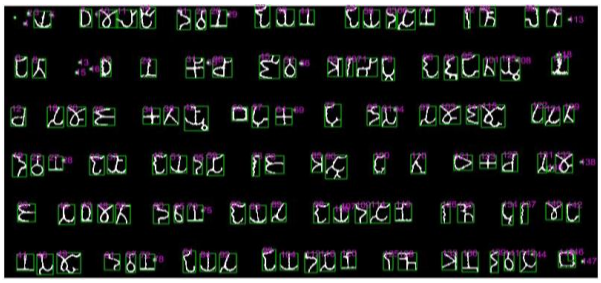

Figure 3(a): Segmentation of characters using bounding box

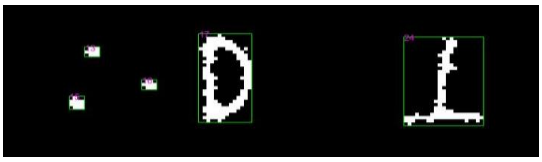

Figure 3(b): Separately segmented dots

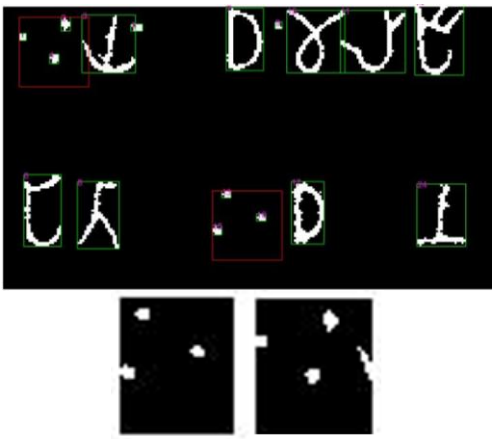

Figure 4(a) and 4(b): Successfully segmented special character

Proposed method is tested on 50 images, containing total 207 "∴" characters and out of all, 189 characters got successfully segmented. 14 wrong objects also got segmented and 18 characters did not segment. Overall percentage of successful segmentation is 91.30%.

4. Conclusion

Considering the results obtained, conclusion derived is, that proposed method is working well for segmentation of "∴"with 91.30% of success. Though there are limitations, the proposed method has scope of improvement.

Acknowledgement

The authors would like to thank Dr. C. H. Patil and Mrs. Rashmi Lad for their support in the research work.

References

1. Naresh Kumar Garg , Lakhwinder Kaur , M.K.Jindal, "Segmentation of handwritten hindi text ",2010 International Journal of Computer Applications(0975-8887) vol 1-No.4
2. M.K.Jindal and R.K.Sharma,G.S.Lehal, "Segmentation of touching characters in upper zone in printed gurmukhi script"
3. Bikash Shaw,Swapan Kuman Parui and Malayappan Shridhar, " A Segmentation based approach to offline handwritten devanagari word recognition" , Internation conference on information technology.
4. Munish kumar,M.k.Jindal and R.K.Sharma,"Segmentation of isolated and touching characters in offline handwritten gurmukhi script recognition ",I.J.Information Technology and Computer Science,2014, 02,58-63, published online January 2014 in MECS.
5. Naresh kumar Garg,Lakhwinder Kaur and M.K.Jindal ,"A New method for line segmentation of handwritten hindi text" ,2010 Seventh internationalconference on information technology.

6. Bikash Shaw,Swapan kr.Parui and Malayappan Shridhar,"Offline handwritten devanagari word recognition: A segmentation based approach"
7. Veena Bansal,R.M.K.Sinha, "Segmentation of touching and fused devanagari characters" ,Pattern recognition 35(2002) 875-893.
8. Saiprakash Palakollu ,Renu Dhir and Rajneesh Rani,"Handwritten hindi text segmentation techniques for lines and characters", Proceedings of the world congress on engineering and computer science 2012 vol I WCES 2012 ,October 24-26-2012,San Francisco,USA.
9. Richard G.Gasey and Eric Lecolinet ,IBM Almaden research center and ENST paris,"Strategies in character segmentation :A Survey".
10. Seong-Whan Lee,Member,IEEE computer society, Dong-June Lee ,Member,IEEE, and Hee-Seon Park,Member ,IEEE,"A New Methodology for Gray-Scale character segmentation and recognition ",IEEE Transaction on pattern analysis and machine intelligence, vol.18,no 10,October 1996.
11. Ashwin S Ramteke,Milindi E Rane ,"Offline handwritten devanagari script segmentation",International journal of scientific & technology research volume 1,issue 4,may 2012.
12. Su Liang, M.Shridhar and M.Ahmadi,"segmentation of touching characters in printed document recognition",Pattern recognition ,volume 27,no.6,pp.825-840,1994.
13. Utpal Garain and Bidyut B.Chaudhuri,Fellow,IEEE "Segmentation of touching characters in printed devnagari and bangala scripts using fuzzy multifactorial analysis",IEEE Transactions on systems,man,and cybernetics-part c : application and reviews, vol.32,no.4,November 2002.
14. Manish Kumar Jindal,Gurpreet Singh Lehal and Rajendra Kuman Sharma,"On segmentation of touching characters and overlapping lines in degraded printed gurmukhi script",International journal of image and graphics vol.9,no.3(2009) 321-353.
15. Richard G.Casey and Eric Lecolinet,Member,IEEE,"A Survey of methods and strategies in character segmentation",IEEE Transaction on pattern analysis and machine intelligence,vol.18,no.7,july 1996.