IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol. 8, Issue.4, Apr 2020 E-ISSN: 2347-2693

Character Segmentation on Degraded Printed ODIA Script

Ipsita Pattnaik[1*], Tushar Patnaik[2]

[1] C-DAC, Noida, India
[2]C-DAC, Noida, India

DOI: https://doi.org/10.26438/ijcse/v8i4.3445 | Available online at: www.ijcseonline.org

Received: 08/Mar/2020, Accepted: 13/Apr/2020, Published: 30/Apr/2020

***Abstract*—** In this paper segmentation procedure of degraded script have been proposed of Odia script. A dataset of 50 documents including 170 words in each document making of 5000 character have been taken after scanning. After that segmentation procedure have been applied to get the accuracy rate of degraded printed Odia script. Also, different level of degradation in a script have been mentioned. Character segmentation on degraded odia printed script have been a tough task due to its Curvy with round format. Due to this style of writing it becomes difficult to segment its Characters. Character Segmentation is an essential part of Optical Character Recognition. Optical Character Recognition is an emerging area of research which helps in converting scanned image or handwritten notes into digital format.

Keywords- Character segmentation , Connected Components, Degraded Script, Optical Character Segmentation, Odia Script

I. INTRODUCTION

Introduction Optical Character Recognition is a common method of digitizing printed text so that they can be electronically edited, searched, store more compactly, displayed on-line. Today reasonably good OCR can be bought for as little as 100$. However, they are only able to recognize high quality printed text documents or neatly written hand printed text. The current research on OCR is now addressing documents that are not well handled by the available system, including several degraded, omni font machine printed text[1]. A variety of approaches have been proposed and tested by researchers in different parts of the world, including statistical methods ,structed and syntactic methods, model matching, neural networks, and expert systems.[2] It consists of various stages pre-processing, classification, Post-Acquisition-Level, Processing, Post-level Processing ,Feature Extraction. This paper gives a procedure for segmenting characters in degraded Odia script. Optical Character Recognition have been a great interest to many computer scientist, engineers and people from other discipline.[3]However Dusshwar and Jha was of the view that accuracy rate of 80% to 90% on neat, clean , hand-printed characters can be achieved by pen putting software.[4] Odia being the regional language of Odisha and second official language of Jharkhand is spoken by 35million people. A very few good result have been obtained in this work. In this paper connected component strategy have been applied to get a better result.

II. OBJECTIVE

To find the accuracy rate of character segmentation of Degraded printed Odia script using Connected Component Strategy.

III. METHODOLOGY

Total 50 documents with 5000 characters have been with 300dpi after that pre-processing procedure have been followed of Grayscale converting and binarization. Binarization method is done by threshold procedure to make it computer readable at the end connected component method is applied to segment the character present in the documents.

IV. FEATURES OF DEGRADED ODIA SCRIPT

It The performance of an OCR system depends upon printing quality of the input documents. Many OCRs have been designed which correctly define fine printed documents in Indian and other scripts. Little reported work have been found on recognition of degraded documents. The performance of any standard OCR system working for fine printed documents decreases, if it is tested on degraded documents. Character Segmentation is an important task for designing an OCR for recognizing degraded documents.

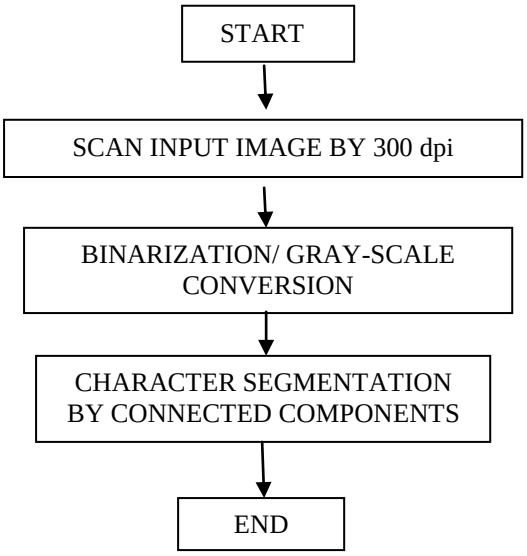
The Odiya script by which Odia language is written, is developed from the Kalinga script, one of the many descendants of the Brahmin script of ancient India. The alphabet of the modern Oriya Script consists of 11 vowels and 41 consonants. These characters are called basic characters and the basic characters of Oriya Script. [5]

In this we discuss the problematic structural features selected for recognizing degraded printed Oriya Script Characters.

- Curvy style formatting makes it difficult to distinguish between its characters.

- Wrong ink density or spread on ink in wrong way on script.
- Discoloration of the script and termites infected documents.
- Align of the documents after printing is also not guaranteed.
- Creases on the documents when folded and unfolded at times
- Not proper ink covering whole character makes it difficult to recognizing
- Uneven spacing between the line and the paragraph also makes it difficult to recognize by the system.

V. OVERVIEW OF THE PROCESS



SCANNED INPUT IMAGE- Scanning a document in 300 dpi (dot per inch) is not considered to be an official standard but considered as a gold standard scanning for Optical Character Recognition (OCR). Scanning at lower dot per inch (dpi) images loses it clarity and quality. Whereas higher resolution will produce bigger image and will cause more time for Image Processing.

BINARIZATION/ GRAYSCALE CONVERSION - RGB image is converted into Grayscale image as less information is provided for each pixel. Whereas the Grayscale conversion is that conversion where all shade of gray are present.

Binarization resolves any image into foreground and background. A binary image is one that consist of pixel that can have exactly two colors, usually black and white. That means pixels are stores in single bit that is 0 and 1. Thresholding method have been applied to get binarization. Where we classify the pixel value. If a pixel value is greater then a threshold value, it is assigned one value (may be white) ,else it is assigned another value (may be black). Function used is cv2.threshold.

- CHARACTER SEGMENTATION BY CONNECTED COMPONENTS- Connected Components refers to a set of pixels having same value connected to each other in a way that there exits a path between every two pixel of the connected component set.[6] Connected Components , in a 2D image , are clusters of pixels with same value . , which are connected to each other through either 4-pixel, or 8-pixel connectivity. 4-pixel connectivity would group all pixels that contact each other or either of there four faces , while 8-pixel would group pixels that are connected along any face or corner.The algorithm used to obtain the connected components is a simple iterative procedure which compares successive scanlines of an image to determine whether black pixels in any pair of scanlines are connected together .[7]

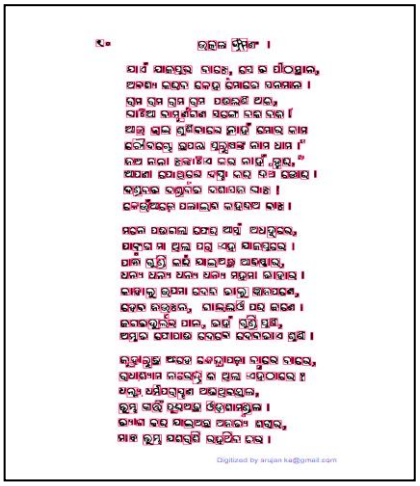
VI. EXPERIMENTAL RESULTS

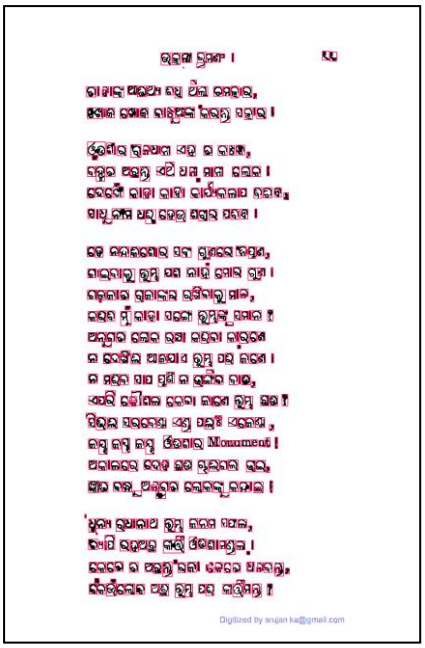
A dataset of 50 degraded document have been taken with a set of 5000 characters was tested including constants, special characters, modifiers, and digits from 0 to 9 . The dataset fonts are random and not of some similar font.. Greater the number of dataset characters, greater is the system efficiency.

The proposed segmented system have been implemented using Spyder version 4.1.2 . The Connected component strategy gives 98.50% accuracy in segmenting the Degraded Odia Script Documents. The accuracy result of 5 document have been show in TABLE I. Whereas the segmentation result of three documents have been mention in FIG.1 .

TABLE I.
Percentage of accuracy of character segmentation

Document	Number Of Characters In single document	Total Character Detect	Percentage Of Accuracy
1	360	358	99.44
2	320	314	98.12
3	420	414	98.57
4	366	360	98.36
5	380	378	99.47





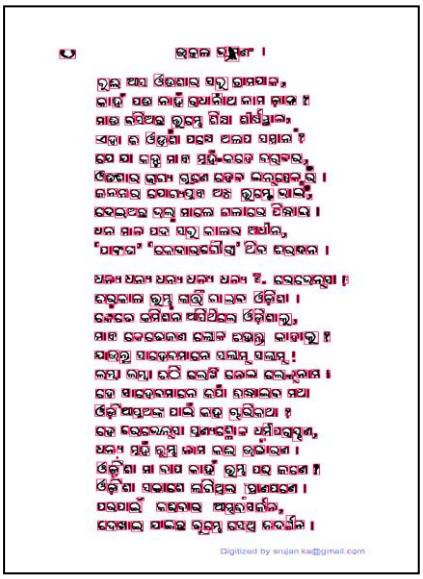


Fig 1. Experimental results of character segmentation

VII. DISCUSSION AND CONCLUSION

After A connected component-based segmentation system have been introduced in this paper for segmenting the characters of 52 Oriya characters. The zero iterations have been applied in connected components for segmenting the characters. As, a result the connected and touching components were segmented by finding the least pixel density at connected characters. This procedure has been worked on several other languages and gave the 98.50% accuracy. Further Vertical Projection profile could be applied to get the segmented characters in a better shape. This would help to segment the Matras and other constraints better. The character segmentation in degraded Odia printed script is been an important factor for recognition in OCR. The connected component strategy described in this paper will find potential application in printed document ,document reading, conversion of any printed document into structural text from and Optical Character Recognition .

REFERENCES

[1] O. D. Trier, A. K. Jain and T. Taxt, "Feature extraction methods for character recognition: – A survey", Pattern Recognition, Vol. 29(4), pp. 641-662, 1996.

[2] C. Y. Suen, "Character Recognition by Computer and Applications", in Handbook of Pattern Recognition and Image Processing, New York: Academic pp. 569-586, 1986.

[3] S. Impedovo, L. Ottaviano and S. Occhinegro, "Optical Character Recognition - A Survey", International Journal of Pattern Recognition & Artificial Intelligence, Vol. 5, pp. 1-24, 1991.

[4] P.Dussawar at. Al.Text Extraction from Complex Color Image Using Optical Character Recognition", Vol.4, pp.730-735, 2015.

[5] U. Pal1, at al., A System for Off-line Oriya Handwritten Character Recognition using Curvature Feature ,IEEE,2007.

[6] Isha Sehgal and K.S. Venkatesh ,"Connected Component Labeling For Binary Images.", International Journal Of Advanced Research, Int.J. of Adv. Res. 7 (8). 916-927.

[7] A. Amin, "Recognition of printed Arabic text based on global features and decision tree learning techniques", Pattern Recognition, Vol. 33, pp. 1309-1323, 2000.

[8] S. Kahan, T. Pavlidis and H. S. Baird, "On the Recognition of Printed Characters of Any Font and Size", IEEE Transactions on PAMI, Vol. 9(2), pp. 274- 288, 1987.

[9] M. K. Jindal, G. S. Lehal and R. K. Sharma, "Segmentation Problems and Solutions in Printed Degraded Gurmukhi Script", International Journal of Signal Processing, Vol. 2(4), pp. 258-267, 2005.

[10] P. D. Gadar, M. Mohamed, and J. H. Chiang, "Handwritten Word Recognition with Character and Inter-Character Neural Networks", IEEE Transactions on Systems, Man, and Cybernetics-Part B:Cybernetics, Vol. 27(1), pp. 158-164, 1997.

Authors Profile

Ms.Ipsita Pattnaik pursed Bachelor of Technology from Guru Gobind Singh Indraprastha University, New Deldi in 2018 . He is currently pursuing M.tech in Computer Science from C-DAC, Noida.

Mr Tushar Patnaik is a Joint Director in C-DAC ,Noida . He has published several papers in natinal , international general . and conferences including IEEE and it's also available online. His main research work focuses on Pattern Recognition,Image Processing, Feature Extraction, Computer Vision, Pattern Classification Digital Image Processing, Machine Learning ,Feature Selection and Signal, Image and Video Processing. He has several years of teaching Research Experience.