Character Segmentation Of Degraded Odia Script

Ipsita Pattnaik*, Tushar Patnaik

Computer Science Department
CDAC, Noida, India

**Corresponding author's email: ipsitapattnaik77 [AT] gmail.com*

ABSTRACT— ***Optical Character Recognition (OCR) is a field which converts printed text into computer understandable format that is editable in nature. Odia is a regional language used in Odisha, West Bengal & Jharkhand. It is used by over forty million people and still counting. With such large dependency on a language makes it important, to preserve its script, get a digital editable version of odia script. We propose a framework that takes computer printed odia script image as an input & gives a computer readable & user editable format of same, which eventually recognizes the characters printed in input image. The system uses various techniques to improve the image & perform Line segmentation followed by word segmentation & finally character segmentation using horizontal & vertical projection profile.***

Keywords— OCR, Odia, Script, computer readable, editable, character segmentation, projection profile

1. INTRODUCTION

In the era of Artificial Intelligence, Optical Character Recognition (OCR) is an important & interesting area solving various real life problems & making work easier. With limitation of not being able to preserve printed paper text, it becomes essentially important to guard a copy of that document that is editable in format. Real life scenarios require some images which are needed in format such as ASCII that can be read, & edited by the user as per specification, to make it convenient our proposed system provides an efficient & time saving way to convert images of a computer printed document into user editable text documents or other formats. It is also referred to as Document Analysis Recognition. The Document Analysis Recognition is of two types: (a) handwritten character recognition, (b) printed character recognition. The applicability of printed character recognition is further segregated into two parts namely: (i) good quality/standard printed documents, (ii) degraded printed documents. There is a massive amount of data in printed odia form that requires digital conversion & that can be reworked upon which could save time & help multiple target audience that requires this data. The printed odia script is scanned for the input image. The system works on scanned input image to make it more informative, beautification strategies are applied. After text & document enhancement & removing constraints from the image the conversion process is done. Initially the input image is converted into grayscale image making it easy for binarization & identifying errors/noises or jitters in the pixels. A bi-map image is formed representing the intensity of each pixel of scanned document. Using techniques of projection profile, we then first divide the document into rows denoting line segmentation, then each line is broken down in parts to perform word segmentation & finally each each word is evaluated to achieve character segmentation.

2. PREVIOUS WORK

Previous work done in this problem domain uses different techniques at different stages. Choksi et al. [1] uses fuzzy KNN to recognise hindi script characters from printed scanned image. It identifies touching characters as a problem. Fuzzy KNN Classifier is paired up with geometric & wavelet transform to solve the issue of touching characters. Pushpalata et al. [2] divides the recognition of Odia characters into: preprocessing, segmentation, feature extraction & classification. Singh et al. [3] analyses various segmentation techniques on Brahmi script. Hossain et el. [4] performs character recognition using neural nets & divided the task into: (a) Preprocessing, (b) segmentation, (c) training recognition & (d) post processing. [5] presents multi-layered feed forward nueral network for handwritten character recognition of English language. Preprocessing in [4] involves rotation, scaling, binarization, noise elimination. [3] uses MATLAB for segmentation.

3. PROPOSED SYTEM

The proposed system receives scanned input image & use techniques to make it more informative & perform segmentation. Fig. 1 depicts the workflow of proposed system. The system is divided into 6 steps: (a) Skew Detection & Correction, (b) Gray Scale Conversion & Binarization, (c) Noise Reduction, (d) Line Segmentation, (e) Word Segmentation, & (f) Character Segmentation.

A. Skew Detection & Correction

We employ strategy of Horizontal projection profile presented in [1]. We update an 1-D array with size equalling number of rows in image. Each value stores number of black pixels in each row of image. We perform rotation & stop when skew becomes zero degree. The stopping condition is that the histogram of image gives maximum amplitude & frequency when skew equals zero degree.

B. Gray Scale Conversion & Binarization

The corresponding rotated image is converted into its grayscale equivalent. Once the image is converted into gray scale format, we now perform binarization. The pixels are stored with intensities ranging from 0 to 255 in grayscale image. We fix the threshold to 127, the pixels with intensities less than threshold are mapped to zero, & pixels with corresponding intensities greater than, equal to 127 are mapped to one. The resulting image is a bi-map image, storing pixels either in 0 or 1. The representation of image is given by black & white color.

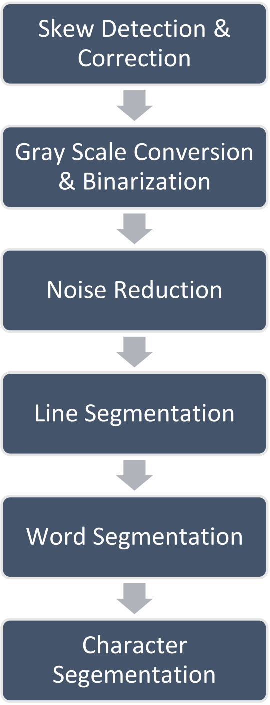

Fig. 1. Proposed System Diagram

C. Noise Reduction

The noise refers to the unwanted signals in the image that degrades its quality. Noise in an image may be introduced while scanning, transmission stages. The presence of noise can exponentially reduce the efficiency & accuracy of the system. It is important to reduce these noise. We use general morphological function which incorporates dilation & erosion methods to remove noise.

D. Line Segmentation

We perform line segmentation using a horizontal sweep line. The document is scanned from top to down. The first black pixel encountered mark the beginning of a new line. The line keeps on sweeping until it finds no black pixel, at this point, it is marked as end of that line. Similarly the document is divided into lines using start-stop pair of sweep line.

E. Word Segmentation

To segment words, we use procedure similar to line segmentation. Here each line is scanned from left to right using a vertical sweep line. In same manner we will employ a pair mechanism, to mark start & end of a word. We introduce a threshold distance d, that determines the general space distance between two consecutive words. In general, space between two consecutive characters is less than space between two consecutive words. We use this notion to segment words. The vertical sweep line scans from left to right. The first black pixel marks the start of the word. The sweep continues till it finds no black pixel for a distance more than d, it marks this point as end of the word. Similarly the line is broken down into words.

F. Character Segmentation

Vertical sweep line is used to segment characters in fashion similar to word segmentation. Here the start of the word is first pixel encountered by vertical sweep line scanning from left to right. The point when no black pixel is encountered, it mark it as end of a character. This process continues till whole document is divided into characters.

4. RESULT COMPARISON & DISCUSSION

We identified different strategies used in previous work & their impact on performance measures. The methods used for binarization, & segmentation incorporates various problems which hamper the performance of the system.

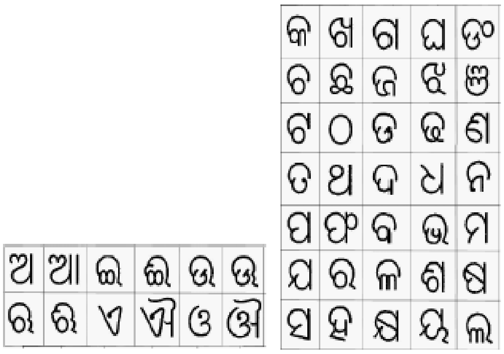

Fig. 1. Odia Vowels & Consonants [2]

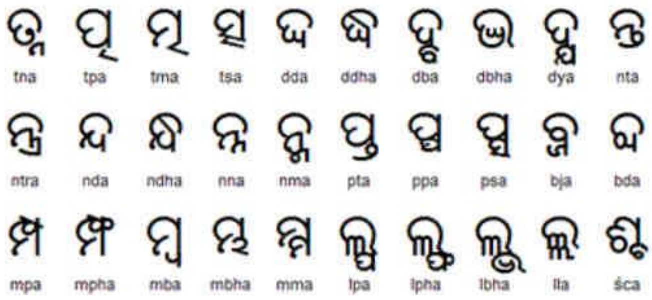

Fig. 2. Common Odia Conjunctions [2]

TABLE I. SUMMARY OF LITERATURE

	Work Details		
	Method Used	***Problems Identified***	***Performance***
[1]	Uses fuzzy KNN with geometric & wavelet transform.	-Joint, touching characters are tough to identify. -Attachment of modifiers within symbol causes problem.	Efficiency: 98.12 %
[3]	Projection Profile for segmentation.	-Overlapping lines leads to mis-segmentation. - Over Segmentation due to break in characters. - Noisy images hampers performance.	NA
[4]	Character Recognition using neural nets.	-Errors due to difficult characters in the Bangla script.	Succes Rate: -Line : 98.8 % for 920 lines, - Word: 96.2% for 10,400 words, & - Characetr: 81% for 10,400 words.
[5]	Multi layered feed forward neural network for character recognition.	-Time consuming to train model. -Complex hidden layers.	Accuracy: 90.19% with two hidden layers.

5. CONCLUSION

We have proposed a system converts printed odia text document into editable digital format using segmentation. We have also employed techniques to remove noise from images & perform skew rotation for better results. We studied various approaches used in previous work & identified problems associated. Our proposed system removes these problem & gives better performance metrics. The number of character set used for training is reasonably low and the accuracy of the network can be increased by taking more training character sets.

6. REFERENCES

[1] Prof. Amit Choksi1 , Kajal Kumari , Shivani Kanojiya , Pragya Sahu, Nishtha Rindani, "Hindi Optical Character Recognition For Printed Documents Using Fuzzy K-Nearest Neighbor Algorithm: A Problem Approach In Character Segmentation", IOSR Journal of VLSI and Signal Processing (IOSR-JVSP), Jan.-Feb. 2018.

[2] Pushpalata Pujari, Babita Majhi, "A survey on odia character recognition", International Journal of Emerging Science and Engineering (IJESE), February 2015.

[3] Ajay P. Singh and Ashwin Kumar Kushwaha, "Analysis of Segmentation Methods for Brahmi Script", DESIDOC Journal of Library & Information Technology, March 2019.

[4] SK Alamgir Hossain, Tamanna Tabassum, "Neural net based complete character recognition scheme for Bangla printed text books", 16th Int'l Conf. Computer and Information Technology, 8-10 March 2014.

[5] J.Pradeep, E.Srinivasan, S.Himavathi, "Neural Network based Handwritten Character Recognition system without feature extraction", International Conference on Computer, Communication and Electrical Technology – ICCCET 2011, 18th & 19th March, 2011.