IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 4, Issue 11, November 2015*

# Study and Analysis for Development of an Efficient OCR for Printed and Handwritten ODIA Documents: A Survey

**Anupama Sahu[1], Sarojananda Mishra[2]**

Department of CSE&A, IGIT Sarang, Dhenkanal, Odisha, India[1, 2]

**Abstract:** The OCR (optical character recognition) is the process of translating the hand written or printed text into a format that is understood by the machine for the purpose of editing, searching and indexing. Preprocessing, segmentation, features extraction, classification and post processing are the main phases of any OCR system and these specific fields are in use today. For all these tasks the segmentation plays a very crucial role in the overall performance of the OCR system. Segmentation can further divided into line, word and character. In this paper, we have discussed different segmentation methods used in various domains. Some of the methods are used for handwritten documents and some of the methods are printed documents. The major focus of this research is to identify the approach that can be segmented into compound and fused character symbols. After the analysis of the existing segmentation methods, we have concluded the favored methods for compound and fused character symbols which are better for the next research. Segmentation is always a frontier area of research in the field of image processing and pattern recognition. There is a large demand for OCR on odia handwritten documents. The objective of this paper is to present a survey of different exiting segmentation methods that have been developed during the last decade. The paper is concluded by suggesting the future aspect of research in this research area.

**Keywords:** OCR, Text line Segmentation, Word Segmentation, Character Segmentation, Odia Handwritten and Printed documents.

## I. INTRODUCTION

Optical character recognition refers to the process of translating or converting the hand written text or printed text into a machine format that is understood by the machines for easily editing and searching. Segmentation of handwritten or printed text into lines, words and characters is one of the major steps in the handwritten text recognition process. A lot of research is done on the printed odia text, but less work has been done on the handwritten odia text recognition. Segmentation of text line by line is an important step because inaccurately segmented of text lines will cause so many errors during the recognition stage. The nature of handwriting makes the process of text line segmentation is the most challenging job. Text of hand written can vary in font, size, color, contrast, alignment and background information. This type of variations is difficult for detect the word in the document. Since handwritten text can vary depending on the user's skill, disposition and way of writing process. Text line segmentation is one of the major components for analysis the image of document. The line segmentation is important information for skew correction, zone segmentation, and recognition of character. The phases of OCR system are shown in the figure below.

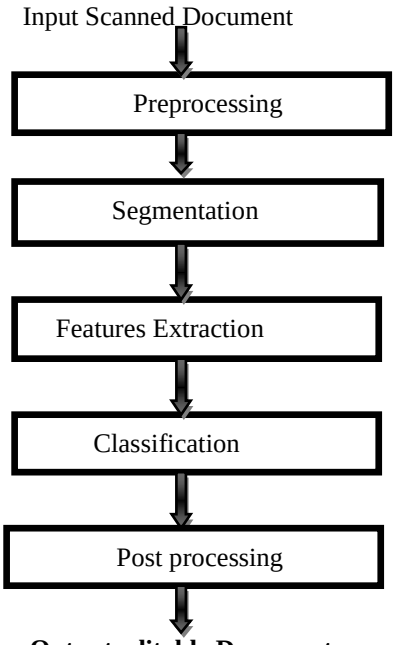


**Figure1: (Phases of OCR)**

Preprocessing:
The raw data goes through some preliminary processing steps to make it usable in stages of character analysis. The main objective of pre-processing is to produce data that can be easily operated for the OCR system. The goals of pre-processing are Binarization, Noise reduction, Stroke width normalization, Skew correction, Slant removal etc.

Segmentation:
The Segmentation divides a text document into line, word and character. Text Line segmentation is based on (Hough Transform, Horizontal projections, smearing), Word segmentation is based on (vertical projections, connected component analysis), Character segmentation is based on (Vertical projections, Feature Extraction) etc.


DOI 10.17148/IJARCCE.2015.41103

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 4, Issue 11, November 2015*

Features Extraction:
The process of extracting the features from object/alphabet to form a feature vector is called feature extraction. With the least amount of elements, the feature extraction can extract a set of features that maximizes the recognition rate.

Classification:
Classification is used to optimized the whole recognition process use several technique i,e k-Nearest Neighbour , Byes Classifier, Neural Networks, Hidden Markov Models (HMM), Support Vector Machines, etc .

Post processing:
The post-Processor is designed to improve the accuracy of the recognition process. The post-processor use different techniques to distinguish one character from another.

This paper provides a survey on segmentation of Odia handwritten or printed documents. This article basically belongs to the field of image processing that designed upon the various analyses upon Odia language and Segmentation of Odia documents. In this research we have reviewed several journals, articles and techniques that are being used for Gurumukhi script, Devanagari scripts etc. Our article levels are as follows. In section II we have provided a literature review upon the various scripts. The Literature review contains the projection based technique, water reservoir principle, stroke and run based features, and neural networks etc that are used in this area of research for various languages. Finally the paper ends up with conclusion.

## II. LITERATURE SURVEY

In this literature review a wide variety of line, word and character segmentation methods for handwritten documents are reported. Some of the methods are used for line segmentation that is projection based method **[1] [2].** In this method the character segmentation regions are determined by using vertical projection profiles and topographic features extracted technique from the gray scale images. Then a nonlinear character segmentation path region is found by using multi stage graph search algorithm. Finally, recognition-based segmentation method is adopted to confirm the nonlinear character segmentation paths and recognition results. Hence it is proved that the proposed methodology is very effective for the segmentation and recognition of overlapped touched characters through the experiments with various kinds of printed documents. In **[3]** they have proposed the technique that, the document image is first captured using a flat bed scanner and pass through different pre processing modules. Then, individual characters are recognized using a combination of stroke and run-number based features. The prototype of the system when tested on a variety of printed material, has achieved 96.3% accuracy in character level.

A water- reservoir principle is proposed in this paper for segmented unconstrained Oriya handwritten text into individual characters. For character segmentation, at first the isolated and touching character in a word are detected. Characters of the word that touch are then segmented using structural, topological and water reservoir concept have achieved the accuracy of 96.7%[4]. The article[5] a neural network is proposed for identification of Gujarati handwritten digits. The classification of digits is suggested by a multi layered feed forward neural network. There are four different profile are used to abstract the features of Gujarati digits. For preprocessing of handwritten numerals the thinning and skew- correction are also used before their classification. For identification of Gujarati handwritten digit, this work has been achieved approximately 82% of success rate. In[6] they have implemented water reservoir based technique for identification and segmentation of touching characters of handwritten Gurumukhi words. Touching characters are segmented based on reservoir area points. We could achieve 93.51% accuracy for character segmentation with this method. The research article[7] presents an algorithm for segmentation of touching Devanagari characters into its constituent symbols and characters. This proposed algorithm extensively uses structural properties of the script of the document. The statistical information about the height and width of character boxes, which are vertically separate from each other, is used to hypothesize character boxes to be touching character boxes. They have achieved the accuracy of 85% when recognition of the segmented touching characters. In[8] Veena Bansal et al. introduced a two pass algorithm for the segmentation and decomposition of Devanagari composite characters or symbols into their constituent symbols. The proposed algorithm basically uses the structural properties of the script. In the first pass algorithm, the words are segmented into separable characters or composite characters. The statistical information about the height and width of each individual box is used to hypothesize whether a character box is composite or not. The hypothesized composite characters are further segmented in the second pass algorithm. The recognition rate of 85% percent has been achieved on the segmented conjuncts**.** In[9] they have presented the Hough Transform technique that used to segment the text line on a subset of the connected components of the document image. Here again the post-processing steps are included for the correction of false alarms and the creation of text lines that the Hough Transform failed to create and vertically efficient separation of connected characters. The distances between adjacent overlapped components in a text line are calculated and the distance is categorized either as an inter-word or an intra-word distance after the comparison with a threshold value. In[10] they have explained the concepts of touching characters and also have presented the survey of various approaches for segmentation of touching character**.** The segmentation of touching Character approaches can be classified into three different types namely, Recognition-free, Recognition-based and hybrid approaches. In[11] they have provide a novel and robust hybrid recognition system for Odia handwritten character . They have designed the standard deviation & zone centroid based feature extraction method for achieving the better accuracy while training and testing the


DOI 10.17148/IJARCCE.2015.41103

Neural Network. The OHCR System is based on the algorithm of feed forward BPNN combined with Genetic algorithm to perform the optimum feature extraction and recognition of character.

## III. CONCLUSION

The study contains a detailed analysis of the different OCR Segmentation technique and their application and also to propose some approaches to combat those techniques. The process of character segmentation has to face many problems like variation in the size of characters, touching characters and overlapping characters. Using structural, topological and water reservoir principle the touching characters of the word are then segmented into isolated characters. In future we want to propose a technique that will work for real time application upon segmentation of overlapping characters. So the character segmentation method can provide a better and improved accuracy in case of Odia Handwriting.

## REFERENCES

[1] B. M. Sagar, G. Shobha, P. Ramakanth kumar, "Character Segmentation Algorithms for Kannada optical character recognition", Proceedings of the International Conference on Wavelet Analysis and Pattern Recognition, 2008.

[2] S. whan Lee, D. June Lee, H. Seon Park, "A new Methodology for Gray-Scale Character Segmentation and Recognition", IEEE transactions on pattern analysis and machine intelligence, 1996.

[3] B. B. Chaudhury, U Pal and M Mitra,"Automatic Recognition of Printed Oriya Script". CVPRU, vol.27, no.1, pp. 23-34, IEEE, 2002.

[4] N. Tripathy and U Pal, "Handwriting Segmentation Of Unconstrained Oriya text". CVPRU, ISI, vol.31[6], pp. 755-769, IEEE, 2006.

[5] A. A. Desai, "Gujarati handwritten numeral optical character reorganization through neural network", Pattern Recognition, ELSEVIER, 2010

[6] M. Kumar, M. K. Jindal, R. K. Sharma," Segmentation of Isolated and Touching Characters in Offline Handwritten Gurumukhi Script Recognition" ,J. Information Technology and Computer Science, 2014.

[7] V. and R. M. K. Sinha, "Segmentation of Touching Characters in Devanagari", Department of Computer Science and Engineering Indian Institute of Technology, Kanpur 208016 India

[8] V. Bansal, R.M.K. Sinha, "Segmentation of touching and fused Devanagari characters", Pattern Recognition, ELSEVIER 2002

[9] G. Louloudis, B. Gato, I. Pratikaki, C. Halatsis, **"**Line And Word Segmentation of Handwritten Documents**"**

[10] A. Kumar, M. Yadav, T. Patnaik, B. Kumar, "A Survey on Touching Character Segmentation ", International Journal of Engineering and Advanced Technology (IJEAT) Volume-2, Issue-3, February 2013

[11] D. Padhi, "Novel Hybrid approach for Odia Handwritten Character Recognition System" Volume 2, Issue 5, Ijarcsse, 2012.