Review on Offline Odia Character Recognition

Deepika Tigga, Sagarika Mishra, Dr. C. S. Panda

Department of Computer Science and Application,
Sambalpur University, Jyoti Vihar, Sambalpur (768019), India

Abstract

Character recognition is an emerging area of research for this current era . Character recognition is a fundamental and challenging field of pattern recognition, machine learning and digital image processing. Various work has been done so far for different Indic language like Hindi, Udru, Gujarati, Punjabi etc. But limited research work has been done over on odia script. So far various techniques have been proposed for character recognition of odia script. This paper tries to review the research work done by various researcher over this ancient script in last few years. This review is divide into three sections. Section I includes introduction of character recognition, odia script and its steps are explained including some features of odia script. Section II includes review literature of various papers and its work in a tabular format. Section III includes conclusion part and future scope.

Keywords: *Optical Character recognition (OCR), pre -processing, Segmentation, Feature extraction, Back Propagation(BP),Neural Network(NN).*

I. Introduction

Character recognition is a electronic means of identifying printed or handwritten documents into machine readable format. Character recognition consist of three key steps:-pre-processing, feature extraction and classification. Character recognition is classified into two classes based on modes of access i.e. online and offline character recognition system[5][7] .In online character recognition system, based on types of recognition is divided into two categories i.e. machine printed and handwritten character recognition . Likewise offline character recognition system , based on types of character recognition is divided into two categories i.e machine printed and handwritten character recognition. Printed characters are uniform and distinctive whereas handwritten are non-uniform and depend on writing style of author [12]. When the term arise offline character recognition ,OCR(optical character recognition) comes into existence. Recognition of character is challenging problem due different font size and different types of variation introduced while writing.

OCR is a process by which handwritten, printed and scanned documents are converted to ASCII character which is recognized by the computer. Generally OCR systems are classified into two types :-

- Template based
- And Feature Based Approach

Template Based:- Template means a system that helps in arrangement of information systematically in computer screen. In template based methodology, an input is unknown character with its pattern ,as each character has its own unique pattern .The input character pattern is matched with the pattern stored. if degree of correlation or similarity is higher than it is assign to its class after classification. Earlier OCR used only template approach but it was not effective due to presence of noise, change in handwriting etc. This method combine with other method can give better result [4][13].

Feature Based:-

In feature based methodology, feature of character is extracted to identify characters based on features that are similar to the features humans use to identify characters. Developers or programmers have to manually determine the properties of characters they feel are important. Some properties as example of feature extraction are Aspect Ratio, pixels above horizontal half point, pixels to right of vertical half point, distance from image centre, reflected y axis , reflected x axis[4][13].

OCR Process

1. **Image Acquisition :-**Image acquisition is a process of capturing the image through camera or some scanner and feeding it to the computer for further processing .The images are represented in the format such as JPEG, BMT, TIF and TNG. The input image may be gray, color or binary tone[4][14].

2**. Preprocessing**:-Pre-processing from its name itself is clear that before processing some work or processing is to be done. Preprocessing consists of series of operation on scanned input image. Operation such as thresholding, noise reduction, binarization , stroke width normalization, skew correction, slant removal, thinning [3][14].

3. **Segmentation:-**Dividing the input image into smaller component called segments. It is done in order to help in extraction of each segment easily i.e. text document into line, line into words and then words into character and then into segments. Segmentation is divided into text line segmentation, word segmentation, character segmentation. In text line segmentation techniques are Hough Transform, Horizontal projection, smearing etc. In word segmentation are based on vertical projection, connection component

analysis. In character segmentation are based upon vertical projection, feature extraction [3].

4. **Feature Extraction :-**Each character or number has its special and distinct characteristic that represent it uniquely. To find set of parameters that uniquely defines the character is called feature extraction. Features are chosen by considering robustness,accuracy , simplicity of detection ,speed of computation, independence of size and fonts and need of classifier design. Types of Feature extraction are :-

(a) Statistical and geometrical features are obtained by computing ratio of statistical and geometrical moments.

(b) Syntactical/Structural features, that represented by stroke, holes, end point,loops or cross-over points.

(c)Hybrid feature:-constitute suitable combination of statistical and structure features. Two important sub-stages of recognition are feature extraction and classification [14].

5. **Classification:-**It uses the feature extracted in previous stage as input to the classifier. The classifier compares the input with the stored feature to assign a class for the input. Classification is divided into method based on statistics, artificial neural network(ANN),kernel and multiple classifier combination[14].

6. **Post processing:-**Objective of post processing is to detect and correct linguistic misspelling in OCR output text . Post processing steps are used to improve the accuracy of OCR character recognition system. Post processing phase can be divided into three groups (a)manual error correction (b)dictionary-based error correction and (c)context –based error correction[14].

Features of Odia script:-

Among different type of Indic language. One of the popular and oldest script is Odia. Odia script has been developed from Kalinga script as descendants from Brahmi script[4]. Odia script founded in stone engraving ,copper plate and palm leaf manuscript[8]. The modern Odia script consist of simple and complex characters .There are 12 vowels,37 simple consonants and 10 numerical digits and near 200 composite character(juktas) in Odia character .Odia script ,by which Odia language is written.The alphabet of the modern Odia script consist of 12 vowels and 35 consonants .These characters called basic characters and Odia Numerals of Odia script .Writing style of scripts is from left to right. The cursive shapes of Odia letters appear to be influenced by Southern Script.Sometimes consonant character combine with consonant to form a new character known as matras are added to consonant[14] .In Odia script a vowel is followed by matras ,depending on vowel its matras is placed at right, left or bottom of consonant.

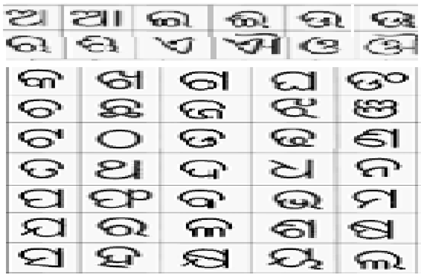


Figure:-odia vowels and consonants

II. REVIEW OF LITERATURE

Basa. D. et. al [1] in review of odia handwritten character recognition in which offline approaches are discussed . The paper discuss about the character modelling ,pre-processing operation required to recognize a text in scanned document . In segmentations handwritten text in divided into Line segmentation are then divided into word segmentation and word segmentation into character segmentation.. Character Recognition is based on method that uses unique structure of some characters has found better result as compared to other methods is discussed throughout the paper .It is concluded with recognition rate is very much affected by similarity of various characters. Similar characters degrade the recognition rate.

Vasudeva.N.et.al[2],represents each single character as a single image ,each containing a single character .Then the gray image format coverts to binary format where each 0 and 1 represents an individual pixel of that image .The binary data is fed into neural network .The output from the neural network is translated into ASCII text and saved as a file . This paper takes the help of Multilayer Perception Neural Network for recognition of offline character for printed text document is used .Difficulty occurs if pre-processing, feature extraction in recognition are done over large volume of data set. It is concluded with reduction of error function results in increase of hidden nodes and epochs for handwritten character recognition.

Sahu.A.et.al[3] focused on the segmented part of character recognition and some methods has been mentioned is used for segment compound and fused character symbol both for printed and handwritten document .Optical handwritten character recognition(OHCR) algorithm is based on forward Back Propagation Neural Network(BPNN) combined with Genetic Algorithm(GA) to perform optimum feature extraction and recognition of character. It is concluded with the problems occurred like size of character, touching character in the word and then isolating the character. So segmentation of overlapping characters in odia handwritten , accuracy can be improved and better.

Panda.S.R.et.al[4] proposed a work on development of algorithm for odia typewritten character recognition using Template with Unicode mapping .A database is created for odia script of pixel size 50*50.In image acquisition the text image that is if RGB then convert it into grayscale..In pre-processing grayscale image is converted to binary image by selecting threshold value is called binarization. Classification is done using template-matching Unicode mapping .Unicode standard has basic principle which emphasizes that each character code width of 16 bits. It is concluded with that the algorithm was successfully tested and achieved accuracy of 97.87% in case of odia characters .Its work is limited to typewritten.

Pithadial.N.J.et.al[5] gives a general discussion of feature extraction technique used in optical character recognition .Offline character recognition has division like magnetic character recognition and optical character recognition. Classification of feature extraction techniques are explained. It is concluded with that efficiency is based applicable. On technique for extraction of features, techniques with different features gives different attributes .

Macwan.J.J.et.al[6] focuses on ascepts of handwritten symbols and problems like rotation problem, scaling problem and shifting problems etc. Water reservoir based technique one of the segmentation technique is used for Odia script. Table V shows the work done on offline odia handwritten Odia script.Major focus is on numerals and less on character. Major challenges and issues are mentioned. It is concluded with the properties of north Indian script further attention is required.

Dedgaonkar.S.G.et.al[7] discussed about several methods like clustering, feature extraction , pattern matching and artificial neural network (ANN) direction based algorithm methods. Various character recognition methods like clustering, feature extraction ,pattern mathching and neural network. It is concluded that template has high speed ,but is ineffective when font slant, font defilement, stroke connection. Combining two or more technique can improve the accuracy rate of the system.

D.Padhi.et.al[8]focuses on the standard deviation and zone centroid based feature training and testing the neural network for recognition of offline character .Work is concentrated on Genetic Algorithm to perform the optimum feature extraction and recognition. It is concluded that method of odia HCR uses a unique and robust combination of ANN and GA gives higher efficiency on programming and testing.

Sk.Md.Obaidullah.et.al[9] deals with development of particular language identification of the script.OCR is discussed its different feature extraction methods are discussed which is required for other scripts and the average accuracy rate of all the script is calculated is shown throughout the paper.

A.Kumar.et.al[10]deals with accuracy and efficiency are parameter of HCR neural network is a technique used to improve accuracy and efficiency .A algorithm is proposed which helps in improving the training efficiency of BP. It is known as Improved BP in which the training processes decreases mean square error(MSE) .It concluded by showing improved BP can speed up convergence of the training process. Improved BP is better than classical BP. Improved BP is effective with smaller values for learning rate.

S.Mishra.et.al[11]deals with OCR technique which depends on the quality of the items been scanned. Proposed a algorithm using NN called feed forward neural network adds unique features of character that helps in feature extraction. It is concluded with the problem of misreading character. The algorithm much depends on quality of items to be processed . It is not objective it does not have the ability to read special character.

Dr. Sarangi.P.K.et.al[12]review on various feature extraction technique. Odia handwritten develop two recognizers transformation scheme i.e DCT and DWT .It has seen that mostly used classifier is ANN with back propagation learning .With DCT and DWT,BPNN is also used to get recognition accuracy upto 92%.It is concluded with that the existing methods for character recognition is not able to perform well in terms of accuracy and efficiency. The existing methods cannot be directly implemented to every language.

Dash.B.et.al [13]proposed a nobel method to recognize offline printed odia character using the concept of DWT. The proposed methodology has five phase .It starts with skew detection and correction of captured text. Then it passes to segmentation and then third phase is to extract the character. And then mapping through template matching. Finally recognizing the character. It is concluded with the method shows good recognition rate upto 90% in offline printed odia character recognition.

Pujari.P.et.al[14]neural network(NN) has been chosen as classifier .Combination of ANN and GA has satisfactory results. Neural Network classifier requires less space and computation time than that of SVM. BPNN based system provides satisfactory performance for Odia recognition than other methods.It is concluded that the curvelet based technique produces highest accuracy and efficiency than any other methods for feature extraction.Every step of odia character recognition is important for system performance.

Tripathy.J.et.al[15], zernike moments feature enhances the recognition capabitlity. Zernike moments is used in the reconstruction process for printed document. It is concluded with development of library for Odia character with different font and size

Mohanty.S.et.al[16] focuses bilingual documents handles both Oriya and Roman script. It calculate accuracy for Odia alphabet with different font and size. It is concluded with selection of features and designing of classifiers jointly lead to better classification performance. Bilingual OCR dealing with degraded, noisy machine printed can be developed further.

Vaidya.S.A.et.al[17] method of feature extraction for handwritten character recognition produces good classification results on handwritten characters. It is concluded Limited to Devanagri and Kannada language can be extended to other language.

Nayak.M.et.al[18] Tesseract is an open source OCR engine. Tesseract can be made to recognize other scripts if the engine can be trained with the appropriate data. It is concluded with Tesseract to recognize the Odia character set, and observed the results.

TABLE I. VARIOUS METHOD'S ACCURACY FROM LITERATURE REVIEW

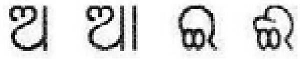
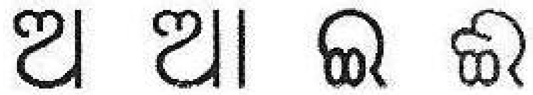
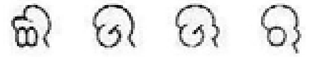
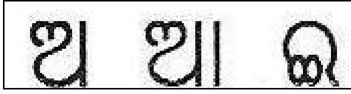
Title of the Paper	Author and Publication Details	Existing Method	Proposed Method	Accuracy rate	Limitation
Offline Character Recognition System using Artificial Neural Network	Nisha Vasudeva et.a l[2] International Journal of Machine Learning and Computing(IJMLC), Vol.2,No.4,August-2012	Bayesian network classifier, Probabilistic methods	Model using mean square error and mean absolute error during ANN training	96%	Efforts to make in getting higher accuracy to improve recognition accuracy.
Odia Offline Typewritten Character Recognition using Template Matching with Unicode Mapping	Smruti Rekha Panda et.al [4] ISACC-2015	Template Matching	Template Matching with Unicode mapping	97.87%	Technique can be extended towards handwritten character and compound character due to its variation in font and size and involvement of different noise and persons.
Novel Hybrid approach for Odia Handwritten Character Recognition System	Debananda Padhi et.al [8] International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE),Vol.2, Issue 5,May 2012	Fractal based texture features, Gabor filter .	Hybrid approach of BPNN and GA	94%	Work to be done on feature extraction of compound character of handwritten Odia text.
Odia Offline Character Recognition using DWT features	Bhabani Dash et.al [13], IOSR Journal of Electronics and Communication Engineering (IOSR-JECE)	Haar wavelet Transform, DCT or DFT	DWT Features extraction method	90%	Limited to printed character and can be extended to handwritten character
A Comparative Analysis of classifiers accuracies for Bilingual Printed Documents (Oriya-English)	Sanghamitra Mohanty et.al [16] International Journal of Computer Science and Information Technologies(IJCSIT), Vol.2,2011	Fractal based texture features, Gabor filter.	K-Nearest Neighbour(KNN), Convolution Neural Network(CNN) and Support Vector Machine(SVM)		Can be extended to handwritten text and more refinement of bilingual OCR.

A Novel Approach of Handwritten Character Recognition using Positional Feature Extraction	Swapnil A. Vaidya et.al [17] , International Journal of Computer Science and Mobile Computing (IJCSMC), Vol.2 ,Issue 6, June 2013	DWT ,zone and distance based feature extraction, LDA based compound distance, Scale Invariant Feature Transform(SIFT)	recognition based on Generalized Regression Neural Network(GRNN)	82.89% in Devanagri, 85.62% in Kannada	Limited to Devanagri and Kannada language can be extended to other language.
Odia Characters Recognition by Training Tesseract OCR Engine	Mamta Nayak et.al[18] , International Conference in Distributed Computing & Internet Technology (ICDCIT),2014	Tesseract version 3.01,adaptive classifier	Procedure to train Tesseract engine for Odia printed text document		Need to train Tesseract for dependent modifier and detect vowel and consonant of odia language.

TABLE II. CLASSIFIER ACCURACIES ON ODIA CHARACTERSE REVIEW

Different types of Classifiers	Accuracy Rate
K- Nearest Neighborhood	***96.47%***
Convolutional Neural Network	***96.53%***
Support Vector Machine	***98.9%***

TABLE III. CLASSIFIER DIFFERENT TYPES OF IMAGES WITH DIFFERENT FONT SIZE 16.2

Image Type	Size of samples	Accuracy percentage
	***Bold and small***	***92.78%***
	***Bold and big***	***98.9%***
	***Normal and small***	***96.98%***
	***Normal and bold***	***97.12***

III. Conclusion

This paper presents a thorough and up-to-date review of Odia character recognition. The research work carried out during the last decade in the field of Odia character recognition has been surveyed. Different approaches employed for each step of Odia Character recognition are also outlined. Each of these methods has its own advantages and limitations. Recognition of character is still a challenging problem since there is a variation in same character due to different font size, different types of noises and involvement of different persons. During the last decades, intensive research studies have been made for recognition of handwritten characters and numerals in various Indian and foreign languages, but a few work has been reported on Odia character recognition. The field of character recognition in Odia language still needs an in depth study. Further Research work is needed to develop Bi-lingual OCR using different font and size of Odia alphabets as well as numerals.

References

[1] D.Basa and S.Meher ,"Handwritten Odia Character Recognition ", National Conference on Recent Advances in Microwave Tubes , Devices and Communication System , Jaipur (March-2011)

[2] N.Vasudeva and Parashar.H.J and Singh .V , "Offline Character Recognition System Using Artificial neural Network",International Journal of Machine Learning and computing(IJMLC),Vol.2,No.4, (August-2012)

[3] A.Sahu and S.,Mishra ," Study and Analysis for Development of a Efficient OCR for Printed and Handwritten ODIA Documents:A Survey" , International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE) Vol.4,Issue 11, (November -2015)

[4] S.R.Panda and J.Tripathy ," Odia Offline Typewritten Character Recognition using Template Matching with Unicode mapping ", ISACC-2015

[5] N.J.Pithadial and Dr. Nimavat.V.D , " A Review on Feature Extraction Techniques for Optical Character Recognition",Vol.3,Issue 2,(February -2015)

[6] J.J.Macwan and M.M.Goswami , "Script Symbol Recognition A Survey on Offline Handwritten North Indian", International Conference on Electrical, Electronics and Optimization Techniques(ICEEOT) (2016)

[7] Dedgaonkar.S.G, Chandavale.A.A , "Survey Method for character recognition " , International Journal of Engineering and Innovative Technology(IJEIT),Volume 1,Issue 5, (May-2012)

[8] Padhi.D , "Novel Hybrid approach for Odia Handwritten Character Recognition System" , International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 5,(May-2012)

[9] Sk.Md.Obaidullah and A.Mondal , "Structural Feature Based Approach for Script Identification from Printed Indian Document" ,International Conference on Signal Processing and Integrated Networks(SPIN),(2014)

[10] A.Kumar and P.K.Bhatia , "Offline Handwritten Character recognition Using Improved Back Propagation Algorithm" ,International Journal of Advances in Engineering Science,(July 2013)

[11] S.Mishra and D.Nanda, "Oriya Character Recognition using Neural Networks" ,(IJCCT)vol.2 Issue 2,3,4 ;(2010)

[12] Dr.P.K.Sarangi, Dr.K.K.Ravulakollu and A.K.Singh ,"Handwritten Character Recognition :A Review", Journal of Management Sciences and Technology(JMST),(October-2015)

[13] B.Dash, S.Pradhan, and D.Rana , "Odia Offline Character Recognition using DWT Features",IOSR Journal of Electronics and Communication(IOSR-JECE)(2016)

[14] P.Pujari, and B.Majhi, "A Survey on Odia Character Recognition",International Journal of Emerging Science and Engineering (IJESE)(February-2015)

[15] J. Tripathy "Recognition of Oriya Alphabets using Zernike moments", International Journal of Computer Application(IJCA), Vol.8,(October-2010)

[16] S.Mohanty, and Das Bebartta.H.N, "A Comparative Analysis of classifiers accuracies for Bilingual Printed Documents(Oriya-English) ",International Journal of Computer Science and Information Technologies (IJCSIT),Vol.2,2011

[17] S.A.Vaidya and B.R.Bombade , "A Novel Approach of Handwritten Character Recognition using Positional Feature Extraction" , International Journal of Computer Science and Mobile Computing (IJCSMC), Vol.2 ,Issue 6, June 2013

[18] M.Nayaka nd A.K.Nayak, "Odia Characters Recognition by Training Tesseract OCR Engine" International Conference in Distributed Computing & Internet Technology (ICDCIT),2014