ISBN: 978-81-923288-4-3

Odia Handwritten Vowel Recognition System by using Discrete Cosine Transform and C5.0 Decision Tree Classifier

Puspalata Pujari[1], Babita Majhi[2]
Department of Computer Science and Information Technology
Guru Ghasidas Vishwavidyalaya, Bilaspur, C.G, India
Email: babita.majhi@gmail.com[1],Pujari.lata@rediffmail.com[2]

***Abstract*—This paper presents a recognition system for Odia vowels by using discrete cosine transform (DCT) and C5.0 decision tree classifier. There are numerous applications of handwritten character recognition system in various fields like banking, document analysis, postal system, form processing, digital library system and many more. Hence a highly robust recognition system is required for such type of applications. In this paper an attempt is made to develop a highly robust system for the recognition of handwritten Odia vowels. The basic steps of character recognition are carried out. The images are preprocessed before recognition. The images are normalized and Canny edge detector method is applied in the preprocessing stage for detection of edges. DCT transform is applied on the images to extract the features. After feature extraction, principal component analysis (PCA) is applied to reduce the dimensionality of features. The reduced features are given to C 5.0, classification and regression tree (CART), quick unbiased efficient statistical tree (QUEST) and chi-square automatic interaction detector (CHAID) decision tree classifiers. The results obtained from all classifiers with test datasets are compared. It is found that a combination of DCT transformation with C5.0 Classifier yields a highest accuracy of 93.05% for the recognition of Odia vowels.**

Keywords— ***Handwritten Vowels Recognition; Canny Edge Detector; Principal Component Analysis; C5.0***

I. INTRODUCTION

Recognition of handwritten vowels has a vital role in present digital world. It has greater importance in the fields like banking, postal system, job application form sorting, document analysis, digital library system and many more. Handwritten vowels recognition system may be on line or off line. In online hand written vowels recognition system the vowels are processed at the time of creation. In offline system the vowels are first generated, stored in a computerized form and used later for processing. The main problem with handwritten vowels recognition system is with the variation found in handwritten vowels. Variation in handwritten vowels is due to different writing styles, the context of the digit, Medias and devices used. The scanned image may be of different size, slant and strokes. There may be local and global deformations: distortions, different writing styles, and thickness variations, wide variety of scales, limited amount of rotation, added noise, occlusion and missing parts. So a sophisticated vowels recognition system is required for correct identification of handwritten vowels. Most of the character recognition system consists of mainly three phases preprocessing, feature extraction and classification as shown in Fig.1. The preprocessing stage involves noise reduction, slant correction, size normalization and thinning. The main objective of preprocessing phase is to enhance the quality of scanned image. The second step is feature extraction. This is the most important phase of handwritten vowel recognition system. The main objective of feature extraction is to extract all the essential characteristics of the image. The selection of appropriate feature extraction method is important factor in achieving high recognition performance. In this paper a feature extraction method based on DCT transformation is applied to Odia handwritten vowels. The third step is the classification process. The main objective of this phase is to identify the character and assign it to correct class. This stage uses the features extracted in the previous stage to identify the characters. In this paper C5.0 classifier is used for the classification of Odia handwritten vowels.

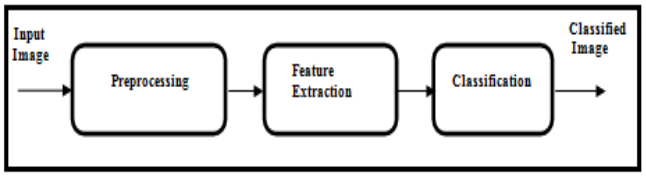

Fig.1. Phases of character recognition system

Oriya officially spelled Odia, is an Indian language belonging to the Indo-Aryan branch of the Indo-European language family. The cursive shape of the Odia letters are influenced from Southern scripts. The curved Shape of the Odia script is a result from the need to write on palm leaves with the help of pointed stylus which prevent to write too many straight lines to avoid damage of palm leaves. Like other scripts Odia script has also 12 vowels shown in Fig.2 below.

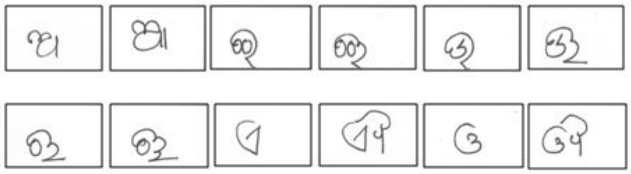

Fig. 2 Odia Vowels

An overview of the paper is as follows: Section II describes a brief review on character recognition with discrete transform with motivation. Section III describes proposed methodology. Section IV describes Canny edge detection techniques used in preprocessing stage. Section V discusses DCT transform used for feature detection and feature reduction. Section VI describes classification phase using C5.0 classifier and tabulates result which is followed by conclusion section.

II. Related Work

Extensive studies have been carried out on recognition of characters in the languages like English, Chinese, Japanese, and Arabic and in many Indian languages like Hindi, Kannada, Tamil, Bangala, Malayalam, Gurumukhi, etc. A few research works has been published in the field of recognition of Odia handwritten vowels . Related work concerns about the previous work and related techniques about character recognition with their results in Odia and other languages. Different approaches used by different authors for feature extraction and classification are as follows

Discrete cosine transform (DCT) has been proposed in [1] to achieve lossless compression of two-dimensional images. Yanping Huang et.al [2] have developed an improved DCT-based method to detect copy-move forgery in images. The method detected duplicated regions even in the presence of blurring or additive white Gaussian noise. A combined classifier model based on two dimensional discrete cosine transform (2D-DCT) and Height Functions (HF) has been proposed in [3] for accurate shape representation and classification. The authors have applied the model on MPEG-7, Kimia-99, Kimia-216, Myth dataset and yielded significant improvement over some well known algorithms. A text dependent writer identification method based on Kannada handwriting has been proposed in [4] by B.V. Dhandra et.al. Radon Transform and Discrete Cosine Transform have been applied for feature extraction. An accuracy of 93.2582% and 100% is achieved by single word and combination of features of three or more words. Discrete Cosine Transform for obtaining feature vectors has been proposed by S. K. Shreedharamurthy et.al [5] for the classification of handwritten kannada numerals. A high degree of accuracy is achieved by this method. In [6] Curvelet transform has been proposed for the recognition of Palamprint. A high recognition rate is achieved by this method. A multifont recognition system using curvelet transform has been proposed by Swati Nigam et.al [7] for the recognition of Odia script. A two stage process has been introduced in [8] for handwritten digit recognition. The authors have used adaptive resonance theory1(ART1)based algorithm in first stage for initial solution and naive evolution strategy in second stage to develop a prototype to reduce the complexity of classification. K- Nearestneighbor classifier has been proposed to achieve an accuracy of 98.73%. Wen et.al [9] presented a Kernel and Bayesian Discriminant based classifier for recognition of Bangla Numeral .They have used characteristics of each class distribution such as the class mean and covariance for classification. A performance of 1.8% error rate is achieved on MNSIT dataset. In [10] Cardoso et.al have represented the data in a feature space using the output of a biologically inspired model. The output of the model has been applied on MNIST and USPS dataset to train a linear classifier for recognition of handwritten digit.

From literature review a little work has been done on Odia character recognition with DCT transformation and C5.0 classifier. Although some work has been done with mulifont Optical Odia character, the authors have ignored handwritten characters [7]. No standard database is used for the recognition purpose. The sample size for training and testing is very less. This paper deals with the Odia handwritten characters by using DCT transform. The database comprising of 1440 samples is collected from NIT Rourekela for the recognition of vowels.

III. Proposed System

For the proposed system 1440 handwritten Odia Vowels are collected, 120 for each vowels in .jpg format. Each sample is categorized into one out of twelve classes ranging from 1-12. These images are resized to 64x64 pixels and are passed through median filtering to smooth the images. In the preprocessing stage edge detection is performed on each image to take boundary of the image using canny edge detection method. After pre processing DCT transform is applied on the images to extract features from the images. Further principal component analysis (PCA) is applied on the extracted features to further reduce the number of features. The reduced features are applied to C5.0 classifier for recognition of vowels. Fig.3. shows a block diagram of the proposed system.

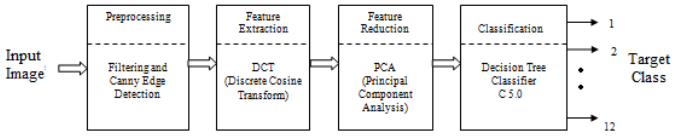

Fig.3. Block Diagram of the proposed vowel recognition system

IV. Edge Detection

Edge detection is the process of finding sharp contrasts in the intensities of an image. It preserves most important structural features and reduces the amount of data in an image. In this paper Canny Edge Detection method is proposed for detecting edges from the images. [11]-[13]. For smoothing, removing noise and unwanted details Gaussian filter $G\sigma(x,y)$ is applied to the image $f(x,y)$ by using equation 1.

$$\left.\begin{array}{c} g(x,y) = G\sigma(x,y) * f(x,y) \\ where\ G\sigma(x,y) is\ given\ as \\ G\sigma(x,y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-\frac{x^2+y^2}{2\sigma^2}] \end{array}\right\} \quad (1)$$