ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol.2, Special Issue 5, October 2014**

# View Based Feature Extraction and Classification Approach to Malayalam Palm Leaf Document Image

Geena K.P[1], Raju. G[2]

Research Scholar, Kannur University, Kerala, India[1]

Associate Professor & Head, Kannur University, Kerala, India[2]

**ABSTRACT:** Malayalam Handwritten Character recognition is still an active area of research. Most of the contemporary works reported use artificial data set taking only 44 characters. Up to 99.78% accuracy is reported with 450 samples per character. In this paper we focus on the recognition of characters extracted from Palm leaf (PL)manuscripts. Unlike the synthetic dataset used PL manuscript images pose more challenges in the pre-processing and recognition stage. To study the performance of the existing Handwritten Character Recognition (HCR) system on PL images, we created a database consisting of 450 samples each of 44 chosen Malayalam character obtained from PL images. For the purpose of comparison a synthetic database consisting of 450 samples each of the same 44 characters are used.

**KEY WORDS:** View Based Feature, Palm Leaf Handwritten Character, MLP classifier.

## I.INTRODUCTION

Handwritten Character recognition has attracted voluminous research in recent times and received extensive attention in academic and production fields. It is an important area in image processing and pattern recognition. India is a multi-lingual and multi-script country, where eighteen official scripts are accepted and have over hundred regional languages. Document Image Processing is one of the key application areas of image processing. Several research works have been focusing toward evolving newer techniques and methods that would reduce the processing time while providing higher recognition accuracy. The recognition of handwriting is complex due to the presence of noise the loss of temporal information. Many promising research results are reported in handwritten character recognition for language like English, Chinese, Korean, Japanese and Arabic. In Indian languages studies are active in Devanagari, Bengali and some promising research findings are also reported in south Indian (Dravidian) languages like Tamil, Telungu, Kannada and Malayalam. The present study is focused on offline handwritten Malayalam isolated characters. Almost all work reported in Malayalam handwritten character recognition was carried out with artificially created data set. Various feature extraction and classification techniques associated with the offline handwriting recognition of the regionalscripts were discussed in the survey. As it is important to identify the script before the recognitionstep, various handwritten scriptidentification techniques were also well-discussed in the survey.A novel approach for recognition of unconstrained handwritten Marathi compound characters was proposed by Sushama&Shaila. The recognition was carried out using multistage feature extraction and classification scheme. The average recognition rate was found to be96.14% and 94.22%, respectively for training and testing samples with wavelet approximation features and 98.68% and 96.23%, respectively for training and testing samples with modifiedwavelet featuresCherguietalpresented an off-line Multiple Classifier System (MCS)for Arabic handwriting recognition which combines two individual recognition systems based on Fuzzy ART network and RBF Some of the recent works reported a recognition accuracy of 99.78%, but used only a subset of Malayalam character set [1][2][3][4][5][6].An automated analysis of palm leaf document images is one of the key research areas. Such an analysis has many applications including creation of text database of historical documents. In this work we carried out a study of the specific challenges posed by a real handwritten character data set extracted from Palm leaf documents. Palm leaf manuscript is one of the oldest mediums of writing in India especially in Southern India. It is also the major source for writing and painting in South and Southeast Asian

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol.2, Special Issue 5, October 2014**

countries. There is no extant of palm-leaf manuscripts in India before the 10th century. However, the palm-leaf was definitely in use much earlier than this since it's mentioned as a writing material in several literary works and its visual representation can be seen in several sculptures and monuments [7][8].

Palm leaf images normally change to blurred images by the presence of noise, low or high contrast, both in the edge area and image area. Pre-processing an image include, removal of noise, edge or boundary enhancement, automatic edge detection, automatic contrast adjustment and segmentation. As multiple noise damages the quality of nature images, improved enhancement technique is required for improving the contrast in palm leaf images [9]. Palm leaf documents are different from other documents that were printed or produced by modern technology. The information on these physical media is harder to extract because the formatting structure of documents are looser. In addition, these documents are of poor quality, due to their fragility and deterioration overage. The various problems are due to issues such as holes and spots on the media, blurriness, smearing, dirt and discoloration. These factors lead to poor contrast, and ghosting noise due to seeping ink from the other side of the manuscripts between the foreground text and the background. In addition, characters were handwritten in narrow spaced lines with overlapping and touching components. Moreover, characters have unusual, varying shapes, and different styles, which depend on the writer. In this paper, an attempt is made to develop a system to recognize Malayalam PalmLeaf characters. Development of an efficient and robust OCR system involves several stages such as pre-processing, feature extraction and classification. However, in this paper suitable technique for extracting the view based features from Malayalam Palm Leaf character which could be further used in development of an OCR system is presented.

Handwritten OCR system for palm leaf document image consists of the following stages:- Image acquisition,Pre-processing, Feature extraction, and Classification and Recognized character. In this paper the performance of view based features in the recognition of Palm leaf character images in carried out.

## II.MATERIALS AND METHODS

### *2.1 Dataset*

For our experiment databases comprising samples of characters from Malayalam palm leaf document image is created. We have collected 4000 Malayalam handwritten palm leaf document images (chadanguBhasha, keralolpathi, krishnagadaha, AdhyathmaRamayanam, Admanandavivekam, Agnihothrachadangu etc.) from the manuscript library of Malayalam Department, Calicut University. The collected images were scanned at 300dpi.The characters are segmented, cropped to fit the minimum sized windows and stored the resultant database consist of 450 samples each of the44 selected character classes. The segmented characters are binarized, resized (72X72) and thinned. Further a synthetic database of same number of samples and characters is obtained.

### 2.2 FEATURE EXTRACTION TECHNIQUES VIEW BASED FEATURES.

Feature extraction is the process to retrieve the most important data from the raw data. The major goal of feature extraction is to extract a set of features, which maximizes the recognition rate with the least amount of elements. In feature extraction stage each character is represented as a feature vector, which becomes its identity [10].View based feature extraction scheme for recognizing Palm Leaf Document image characters is proposed in this work. Everycharacter image of size 72x 72pixels. The view is a set of points that plot one of four projections of the object (top, bottom, left and right) it consists of pixels belonging to the contour of the character and having extreme values of one of its coordinates. For example, the top view of a letter is a set of points having maximal y coordinate for a given x coordinate. Next, characteristic points are marked out on the surface of each view to describe the shape of that view (Figure 2) .The method of selecting these points and their number may vary and can be decided on experiment bases. In the considered examples, twelve uniformly distributed characteristic points are taken for each view.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol.2, Special Issue 5, October 2014**

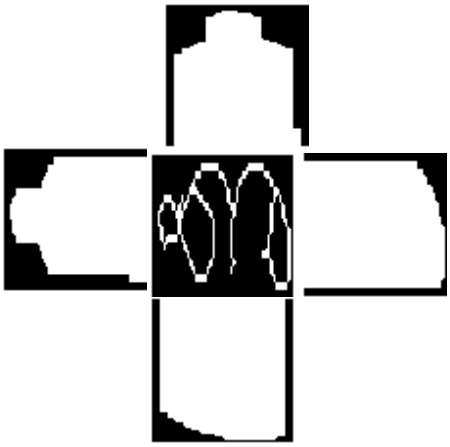

Figure 2.Selecting characteristic points for four views.

**2.4 Feature Extraction Algorithms**

For each image in the database, apply the following method. Find and record the feature values.
Step1: Resize character image to 72X72.
Step2: Find the Left image view at each pixel.
Step3: Find the right image view at each pixel.
Step4: Find the top image view at each pixel.
Step5: Find the bottom image view at each pixel.
Step6: For each view, find the sum of each view code.

Extracted features (12, 24, and 48) are stored into files which are used for subsequent Experiments. The next step is calculating the y coordinates for the points on the top and down views, and x coordinates for the points on left and right views. These quantities are normalized so that their values are in the range [0, 1]. Now, from 48 obtained values the characteristic vector is created to describe the given character, and which is the base for further analysis and classification.

## III.CLASSIFICATION

Back propagation is a popular and widely used network learning algorithm. A back propagation network is used as classifier. A back propagation network is fully connected layered, feed forward neural network containing one input layer with a number of neurons, or more intermediate layers called hidden layers and an output layer .The recognition performance of back propagation network will highly depend on the structure of the network and training algorithm. The number of nodes in input, hidden and output layers will determine the network structure [11]. All the neurons of one layer on fully interconnected with all neurons of the subsequent layer.

## IV. RESULTS AND CONCLUSIONS

At present, we have considered neatly palm leaf characters for the experimentation purpose. The proposed model is implemented using Mat lab in Windows7 platform. After pre-processing,for extracting the features, a total of 19800 samples are used for recognition purpose. The classification is performed using MLP classifier in weka.The results obtained is tabulated in (table1). For both databases, the 75% of each class is used for training while the rest 25% is used for testing. As it in conduct overall recognition accuracy of Handwritten Palm leaf document is lesser than that of synthetic data.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol.2, Special Issue 5, October 2014**

`Table1.Classification Performance of two databases

| Database | Feature Dimension | Accuracy (%) |
|---|---|---|
| **HWPLD** | 12 | 89.66 |
| | 24 | 90.6 |
| | **48** | **94.09** |
| **HWSC** | 12 | 90.07 |
| | 24 | 93.07 |
| | **48** | **95.97** |

In this paper we have carried out a comparative analysis of the performance of view based feature and MLP in the recognition of characters obtained from Palm leaf document images and that of synthetic character images. It in established that the recognition rate of characters obtained from palm leaf document is lesser than that of synthetic data. Hence better HCR system need to be designed for palm leaf character images.

## REFERENCES

1. G. Raju**,** Bindu S. Moni and **Madhu S. Nair**, "A Novel Handwritten Character Recognition System using Gradient Based Features and Run Length Count", **Sadhana**, The Indian Academy of Sciences, Springer-Verlag. (in press).
2. GeenaK.P,RajuG,"Character Recognition System For palm Leaf Document Images", Kerala Science Congress 2014.
3. Chacko, B.P. Babu A.P (2010), Pre and Post Processing Approaches in Edge Detection for Character Recognition.Frontiers in Handwriting Recognition (ICFHR).
4. N.Valliammal,S.N.Geethalakshmi (2011) ,Hybrid Method for Enhancement of Plant Leaf Recognition World of Computer Science and Information Technology Journal (WCSIT) ISSN: 2221-0741 Vol. 1, No. 9.
5. Bindu S Moni and Raju G (2011),a Modified quadratic classifier and directional features for handwritten Malayalam character recognition.Int. J. Comput. Appl. Spec. Issue Computer Science.
6. Jomy John, Pramod K. V, KannanBalakrishnan,"Handwritten Character Recognition of South Indian Scripts: A Review",National Conference on Indian Language Computing, Kochi, Feb 19-20, 2011.
7. OlarikSurinta and RapeepornChamchong(2008), Image segmentation of historical handwriting from palm leaf manuscript 370-375.
8. WafaBousallana, AbderrazhakZahourAdelAlimi(2008),A methodology for the separation of Foreground/Background in Arabic Historical Manuscript using Hybrid Method.Jounal Universal computer science, vol.14, no.2
9. Ntogas,Nikolas,VentZas,Dimirios(2008),binarization algorithm for Historical manuscript,12$^{t}$ WSEAS International conference on Communications, heraklion, Greece.
10. G.Vamvakas, B. Gatos and S. J. Perantonis (2009), A Novel Feature Extraction and Classification Methodology for the Recognition of Historical Documents, 10th International Conference on Document Analysis and Recognition.