iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** VII **Month of publication:** July 2020

DOI: https://doi.org/10.22214/ijraset.2020.30618

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VII July 2020- Available at www.ijraset.com

HOG and Zone Base Features for Handwritten Malayalam Character Classification

Pooja M S[1], Samitha T[2]
[1]*M-Tech student, Mahaguru Institute of Technology, Kattachira, Kerala*
[2]*Assistant Professor, Department of Electronics and Communication, Mahaguru Institute of Technology, Kattachira, Kerala*

***Abstract**: The recognition of handwritten characters remains one of challenging task in character recognition problems. The variations created by each person in writing the characters affect the character recognition result. Many studies have been performed to increase the performance of Malayalam character recognition. The efforts are to extract the best feature for classification or to get the best classifier for classification. In this study, HOG feature and zoning based feature is be used to classify Malayalam Characters. The performance of both features will be compared for classifying Malayalam character by using SVM classifier. The result showed that HOG feature is able to show higher accuracy as compared to the simple zone based feature. The best accuracy for HOG is achieved by using binary input. On the other hand, despite its simplicity zone based feature is able to achieve accuracy by using skeleton input. Considering that the zone based feature used in this research is simply the pixel count in each zone, future research may be performed to extract more statistical properties on each zone. Future works may also focus on rotation free feature extraction for Malayalam character classification.*
***Keywords**: Malayalam character recognition, HOG, Zone based feature*

I. INTRODUCTION

Automatic recognition of handwritten text has been a frontier area of research for the past few decades. The recognition of handwritten characters remains one of challenging task in character recognition problems. The variations created by each person in writing the characters affect the character recognition result. Malayalam handwritten character recognition precision is still inhibited around 70% due to the challenges in Malayalam character set.

The presence of two different scripts old and new script, huge character set, presence of similar shaped characters makes Malayalam handwritten character recognition more difficult. Among the Indian languages, the research on South Indian languages such as Kannada, Tamil, Telugu and Malayalam demands far more attention. The researches on Malayalam scripts have gained high attention in the recent years due to the initiatives of the Government of India, Ministry of Communications and Information Technology and Technology Development of Indian Languages (TDIL).

Effort to improve the performance of handwritten Malayalam character recognitions are still challenge for researcher. Based on the previous works on other language's character recognition, HOG is one of features, which provide good result for character recognition.

On the other hand, Zoning based feature has also become one of good feature to distinct one character to another. In this study, HOG feature and Zoning Based Feature will be used to classify Malayalam Characters. The performance of both features will be compared. The classifier used for this study is SVM which performs well with HOG classifier.

Malayalam is one among the 22 scheduled languages of India. Malayalam is also spoken in the Union territories of Lakshadweep and Mahe. Malayalam script is derived from the Grantham script which is an inheritor of the old Brahmin script. It is one among the 4 major Dravidian languages of South India. Malayalam has close affinity to Tamil.

The Malayalam language was written in Vatteluttu, a script that had evolved from Tamil-Brahmin. It first appeared in Vazhappalli inscription.

The complete character set includes 15 vowels, 36 consonants, 5 chillu, 3 consonant signs, 9 vowel signs, anuswaram, visargam, chandrakkala and 57 conjunct consonants. It also includes 9 numerals which are seldom used. The basic Malayalam character set is represented in Figure 1. During the past few years, many works have been reported in Offline character recognition of Malayalam.

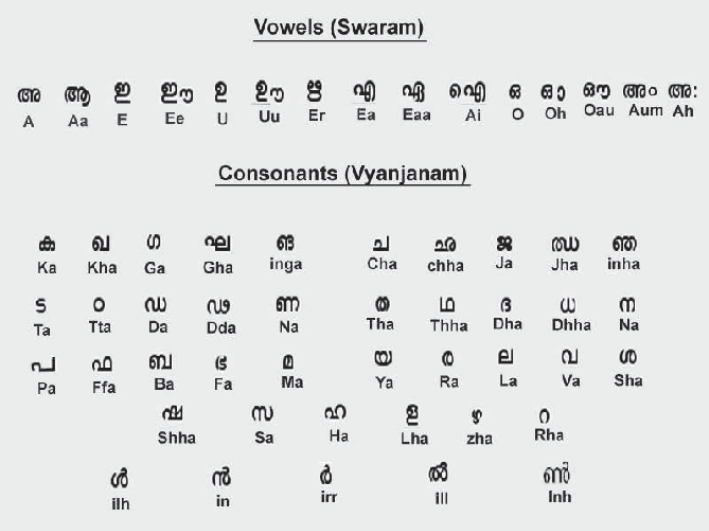


Fig. 1: Basic Malayalam character set

II. OBJECTIVES

The main objective is to identify handwritten isolated Malayalam character with the use of HOG and zone based features and also to develop an algorithm to segment a character of the handwritten word. The existing optical character recognition systems are not suitable for recognizing handwritten Malayalam characters.

III.RESEARCH METHODOLOGY

The proposed system is developed for isolated Malayalam handwritten character recognition. This work uses MATLAB language to implement the OCR system for Malayalam. Malayalam character images are given as input to the system and it recognizes the character and gives the interpreted Unicode to the user. Fig. 2 shows the flow of the research methodology. Image acquisition is performed at the initial stage of the research. Image is scanned from Malayalam character writing. After being scanned, each character is being processed in pre-processing step. The output of pre-processing step is image that is ready for further analysis. The analysis performed in this study are extracting HOG feature, and extracting Zoning based Feature. The input of feature extraction processes is the pre-processed binary image. The output of the analysis process is feature, which will be classified by SVM classifier.

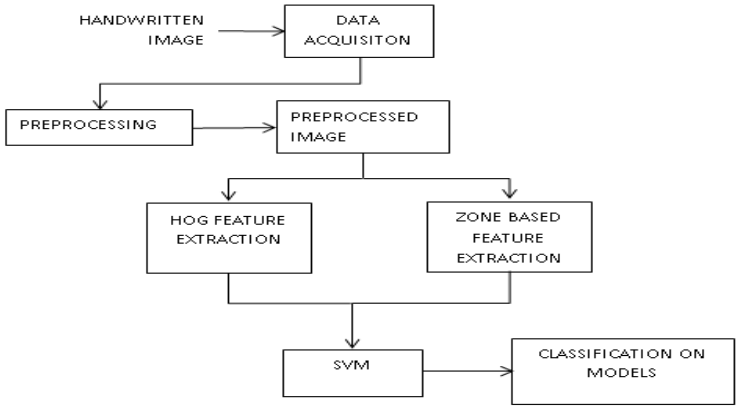


Fig. 2: Research flow

A. Data Acquisition

Malayalam is the official language of Kerala and is spoken by around 35 million people in the world. It first appeared in Vazhappalli inscription. The complete character set includes 15 vowels, 36 consonants, 5 chillu, 3 consonant signs, 9 vowel signs, anuswaram, visargam, chandrakkala and 57 conjunct consonants. It also includes 9 numerals which are seldom used. However, in this research, the classification is performed based on the 33 Basic Malayalam character only.

B. Pre-Processing

The objective of pre-processing phase is to eliminate as much distortions as possible from the scanned image. These distortions occur due to the poor quality of scanners and degraded documents. This approach separates the foreground and background pixels. Upon the removal of the unnecessary white space, image is converted to binary image. Binary image acts as input for the feature extraction process.

C. Histogram of Oriented Gradient

Histogram of Oriented Gradient, or HOG, is one of image descriptor which shows the occurrences of gradient orientation in localized segment of the image. In order to calculate HOG feature, initially the gradient of each pixel in the image is calculated. For every pixel, the gradient consists of magnitude and direction information. The gradient of each pixel is calculated. The histogram of gradient orientation in equally sized cells is then calculated based on the pixel in the corresponding cells. Followings are step by step calculation to compute HOG feature.

1) *Pre-Processing*: This step is to prepare the part of image to be processed. For example if the image to be processed contains a lot of objects, pre-processing steps is performed to crop patch of images with single objects. HOG feature descriptor used for pedestrian detection is calculated on a 64×128 patch of an image. Of course, an image may be of any size. Typically patches at multiple scales are analysed at many image locations. The only constraint is that the patches being analysed have a fixed aspect ratio. In this case, the patches need to have an aspect ratio of 1:2. For example, they can be 100×200, 128×256, or 1000×2000 but not 101×205.
2) *Calculate the Gradient Images:* To calculate a HOG descriptor, first calculate the horizontal and vertical gradients; after all, calculate the histogram of gradients.Gradient of each pixel is calculated by filtering the image with the kernels as shown in Fig 3.

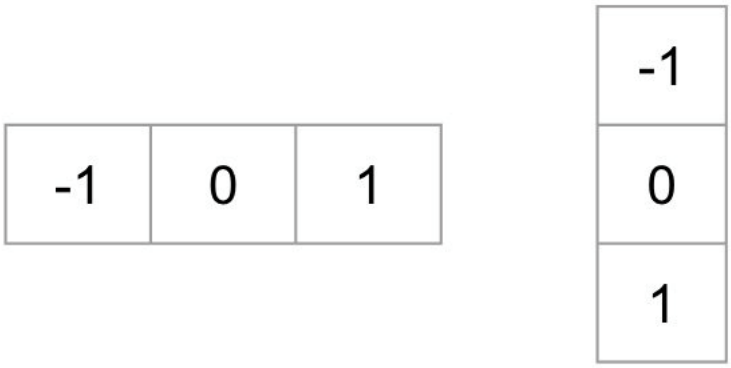


Fig. 3: Kernel to detect gradient of images

3) *Calculate Histogram of Gradients in the Specified Cell Size:* After calculating the gradient, analysis is then perform in an image patch. In this step, the cell size is usually determined prior analysis or by trying different size value and evaluating the accuracy. The histogram for each cell consists of 9 directions. For example, if image size is NxN, and cell size is mxm, total number of cell is (N/m) x (N/m). For each cell, the number of feature is 9.
4) *Block Normalization:* Normalization is performed to normalize the resulting vector. In this case, normalization is done by dividing each histogram with the total length of 4 adjacent cells. The number of feature resulted from this step is 36x1 for 4 adjacent cells.
5) *Calculate the HOG Feature:* The last step is to calculate HOG feature for all area of the image. For example if the input image size is 128 x 128 and the cell size is 8x8, in total there will be 16x16 cells. Each cell will have 9 features. The normalization will introduce new window size of 15x15, because creating sliding window with the size of 4 adjacent cells. Each window will have 36 features. Hence the resulting feature is 36x15x15, or 8100. For cell size of 16x16, the number of cell is 8x8. The total number of normalization window is 7x7, each of which consists of 36 features. Hence, total feature is 7x7x36 which is equal to 1764. Calculate the HOG feature.

D. Zone Based Feature

Zone based feature is a statistical feature which able to represent important information with low complexity and high speed. The features are calculated by first dividing NxN image into non-overlapping cell with equal size. In each cell, the number of pixel is calculated. If the cell size is mxm, the resulting feature size is equal to the number of cell, which is (N/m) x (Nxm).

E. Support Vector Machine Classifier

Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for both classification and regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, plot each data item as a point in n-dimensional space (where n is number of features) with the value of each feature being the value of a particular coordinate. Then, perform classification by finding the hyper-plane that differentiates the two classes very well. Support Vectors are simply the co-ordinates of individual observation. The SVM classifier is a frontier which best segregates the two classes (hyper-plane/ line).

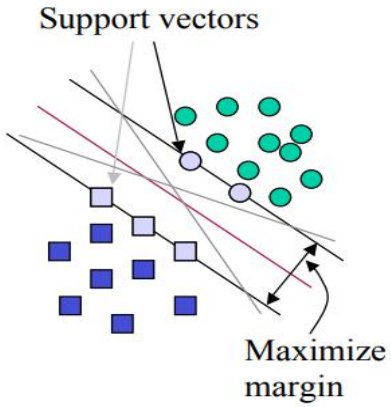


Fig. 4: SVM separating plane

IV. RESULTS AND DISCUSSION

The proposed system is implemented for 33 selected isolated Malayalam characters. The implementation was done using MATLAB R2019b. MATLAB is a high-level language and interactive environment for numerical computation, visualization, and programming. We can analyze data, develop algorithms, and create models and applications using MATLAB. The language, tools, and built-in math functions enables to explore. The language, tools, and built-in math functions enables to explore multiple approaches and reach a solution faster than with spreadsheets or traditional programming languages, such as C/C++ or Java.

A. Pre-Processing

The scanned image is converted to binary image; this approach separates the foreground and background pixels. Then the line segmentation and character segmentation of the collected samples are carried out using horizontal and vertical projection profiles. Fig 5.1 shows the segmentation operation using these methods. Further, the characters are cropped by placing a bounding box around it.

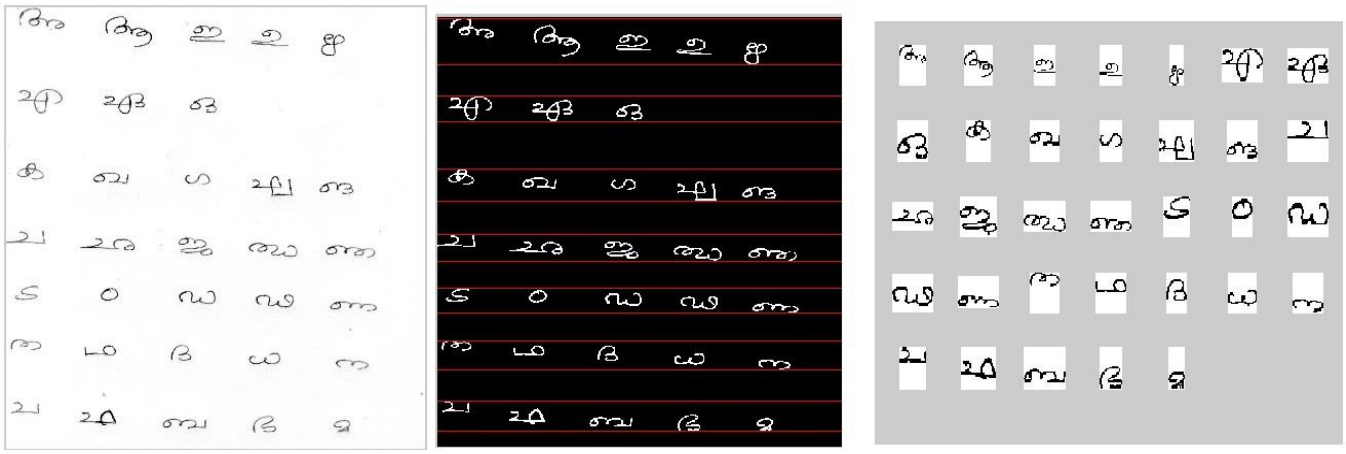


Fig. 5: Line and Character segmentation

B. Feature Extraction and Classification

The most important part in classifying an object is achieving the best feature for classification. Upon the pre-processing step, the feature extraction is performed by extracting the HOG value for all the characters. The cell size to extract HOG features is varied between 8x8 to 64x64. The HOG features serve as input for SVM classifier.

Here, 10-fold cross validation is used to measure the accuracy of the classifier. The results of the experiments are shown in Table.1. It is shown in the table that the highest accuracy is achieved by setting the cell size to be 16x16. In addition to HOG feature, zone base feature is also extracted in this research. Zone based feature for binary image is extracted. For each zone, the number of pixels is counted. The cell size defining each zone is varied between 8x8 pixels to 32x32 pixels. The result of experiment is summarizing in Table 2.

Table 1 Experiment Result Using HOG Feature

HOG Size	Accuracy (%)
8	81.31%
16	83.83%
32	80.80%

Table 2 Experiment Result Using Zone Based Feature

HOG Size	Accuracy (%)
8	69.19%
16	68.18%
32	41.41%

C. Classification

Classification is the final phase in character recognition that assigns a unique label to character images based on the features extracted. The Malayalam handwritten character recognition system proposed here uses binary SVM for classification. The Training images are used to train the machine in order to create a model. Such a model will be used to classify the unlabelled test image. Feature vector is extracted for each of the training images. This feature vector is stored in matrix format. Feature matrix and manually created labels are used to construct a training model.

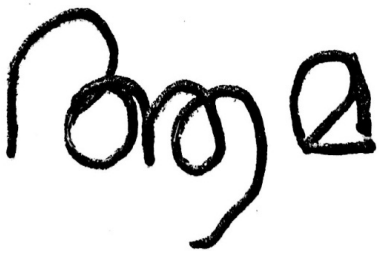


Fig. 6: Test sample

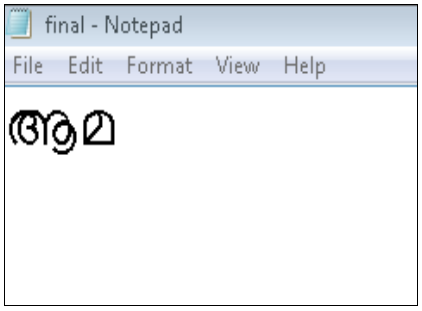


Fig. 7 Recognition result

V. CONCLUSION AND FUTURE WORKS

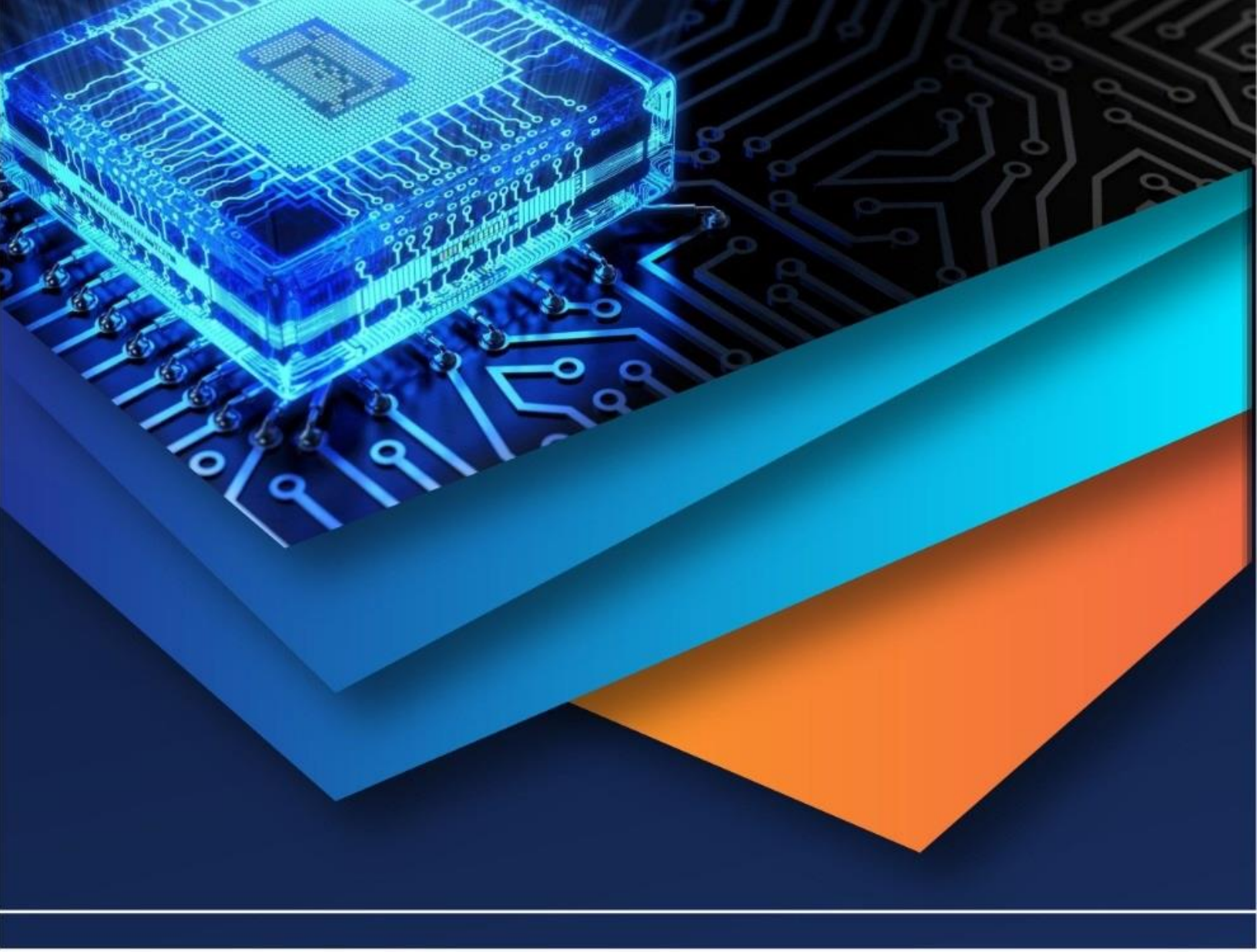
In this research, handwritten Malayalam character recognition is performed by extracting HOG feature and Zone base feature of the character. Inputs of the feature extraction process are binary image of the character. The features are classified by using SVM classifier. The experiment showed that HOG feature is able to show higher accuracy as compared to the simple zone based feature. The best accuracy of 83.83% is achieved by using HOG features. Considering that the zone based feature used in this research is simply the pixel count in each zone. The current work has been proposed for basic Malayalam characters. It can be extended for the recognition of complete Malayalam characters including conjunct consonants as well. The word recognition in Malayalam scripts is also an unexplored area. Most of the misclassifications are due to confusing pair of characters. So the future work can be directed at reducing these errors by incorporating additional features in the post processing stage for identifying these confusing pair of characters.

VI. ACKNOWLEDGEMENT

The authors would like to thank everyone for having given us this opportunity to conduct this study. We would also thankful to all the teaching and non-teaching of Mahaguru Institute of Technology (Formerly, Sri Vellappally Natesan College of Engineering) for supporting us throughout this work

REFERENCES

[1] Panyam Narahari Sastry, T.R. Vijaya Lakshmi, N.V. Koteswara Rao, T.V. Rajinikanth, Abdul Wahab, "Telugu handwritten character recognition using zoning features", (pp. 1-4), IEEE, 2014.
[2] Anitha Mary Chacko and Dhanya P.M.(2014) "Handwritten Character Recognition in Malayalam Scripts-A Review", *Int. Journal of Artificial Intelligence & Applications* (IJAIA), Vol. 5, No. 1, January 2014, pp. 79-89
[3] Rajashekaradhya S. V, vanaja Ranjan P, Manjunath V N, "Isolated handwritten Kannada and Tamil numeral recognition: a novel approach", First International Conference on Emerging Trends in Engineering and Technology 2008 IEEE.
[4] Giridharan. R, Vellingiriraj. E. K, Dr. Balasubramanie. P "Identification of Tamil ancient characters and information retrieval from temple epigraphy using image zoning ", Recent Trends in Information Technology (ICRTIT), 2016 International Conference on(pp. 1-7). IEEE, 2016.
[5] Abdul Rahiman M, M S Rajasree Masha N, Rema M , Meenakshi R, Manoj Kumar G,"Recognition of Handwritten Malayalam Characters using Vertical Horizontal Line Positional Analyzer Algorithm",2011 International Conference IEEE
[6] G. Raju and D. S. Guru, "1D Wavelet transform of projection profiles for isolated handwritten character recognition, Proc. of ICCIMA07, Sivakasi, 2007, 481-485, Dec 13-15.
[7] R. John, G. Raju and D. S. Guru, "1D Wavelet transform of projection profiles for isolated handwritten character recognition, Proc. of ICCIMA07, Sivakasi, 2007, 481-485, Dec 13-15.
[8] Binu P. Chacko, Babu Anto P, "Discrete Curve Evolution Based Skeleton Pruning for Character Recognition, Seventh International Conference on Advances in Pattern Recognition, 2009.
[9] M. A. Rahiman et. al., "Isolated handwritten Malayalam character recognition using HLH intensity patterns, 2010 Second International Conference on Machine Learning and Computing.
[10] Sutha, J. AND Ramaraj, N, "Neural network based offline Tamil handwritten character recognition system", In Proceedings of ICCIMA07,2007 446450.
[11] Bindu S Moni, G Raju,"Modified Quadratic Classifier for Handwritten Malayalam Character Recognition using Run length Count", In International Conference IEEE, 2011.
[12] Vidya V, Indhu T R, Bhadran V K,R Ravindra Kumar, "Malayalam Offline Hand- written Recognition using Probabilistic Simplified Fuzzy ARTMAP". Advances in Intelligent Systems and Computing Volume 182, 2013, pp 273-283.
[13] Jomy John, Kannan Balakrishnan, Pramod K. V, "A System for Offline Recognition of Handwritten Characters in Malayalam Script", I.J. Image, Graphics and Signal Processing, 2013, 4, pp 53-59.
[14] Kumar, S. Sachin and Manjusha, K. and Soman, K.P , " c", Recent Advances in Intelligent Informatics, Advances in Intelligent Systems and Computing Volume 235, 2014, pp 435-442.
[15] Vidya V, Indhu T R, Bhadran V K,R Ravindra Kumar, "Malayalam Offline Hand- written Recognition using Probabilistic Simplified Fuzzy ARTMAP",Advances in Intelligent Systems and Computing Volume 182, 2013, pp 273-283.
[16] Gandhi, R. I. and Iyakutti, K," An attempt to recognize handwritten Tamil character using Kohonen SOM", Int. J. Advanced Network. Appl. 1, 3, 2009, pp-188192.
[17] Hewavitharana, S. and Fernando, H. C." A two-stage classification approach to Tamil handwriting recognition", Proceedings of the Tamil Internet Conference, 2002,pp 118124.
[18] N. Shanthi and K. Duraiswamy,"Performance Comparison of Different Image Sizes for Recognizing Unconstrained Handwritten Tamil Characters using SVM, Journal of Computer Science 3 (9): 760-764, 2007 ISSN 1549-3636.
[19] Sitamahalakshmi, T., Babu, V., and Jagadeesh, M, "Character recognition using Dumpster Shafer theory combining different distance measurement methods". Int.J. Engin. Sci. Technol. 2, 5,2010,pp 11771184
[20] Trier.O.D, Jain.A.K and Taxt.J. (1992), "Feature extraction methods for character recognition - A survey", Pattern Recognition, vol.29, no.4, pp.641-662.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY