www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3 Issue 6 June, 2014 Page No. 6402-6404

Proposed Algorithm for HCRUsing K-Means Clustering Algorithm

Meha Mathur[1], Anil Saroliya[2] Varun Sharma[3]

Amity School of Engineering & Technology
Amity University Rajasthan, India

e-mail[1]: mathur.meha05@gmail.com
e-mail[2]: eranilsaroliya@gmail.com
e-mail[3]: mtechvarun@gmail.com

Abstract- India is a multi-linguisticcountry and Hindi is a national language of India. There is no such work has been done on offline recognition of Hindi characters so that the Hindi data is stored digitally and the paper work will reduce and the data also store safely for the long period of time because as we know that the data on the paper is not secure, paper may lost or may get faded. Therefore in this paper an algorithm is proposed torecognize the Hindi character optically using k-means clustering algorithm. HCR is not same as the English character recognition because Hindi characters are joined together with the shirorekha which is the line on the upper part of the characters and in English language there is no shirorekha. So in English there is no need to remove that shirorekha but for recognize Hindi character it is necessary.
K-means clustering algorithm is used for cluster the same data into their respective clusters and for classification.The objective of this paper is to provide a high performance OCR solution for Devanagari script that can help in exploring future applications such as navigation, for ex. traffic sign recognition in foreign lands etc.
Keywords- OCR, HCR, Hindi, Shrirorekha, Pre-processinig, Segmentation, Feature Vector, Classification, Devnagari.

I. INTRODUCTION

We recognize the problem of Hindi character recognition and propose a mechanism for recognition based on k-means clustering algorithm. HCR is not same as the English as it is having the shirorekha on the top of the words. So we have to divide the letters from the word by removing that shirorekha. The paper introduce propose a two masks one is for horizontal projection and other for vertical projection of gray scale image to detect & eliminate shirorekha of word to decompose into individual characters from the words. Once characters are cropped from the word k means cluster is implemented for converting the image into feature vectors. Feature vectors are primary need for the classification state. Classification state implement elucidian distance method for better classification of the test data from the train data.

To recognize the hindi character from the digital image, image segmentation is to be performed in computer vision. The automatic detection and recognition of hindi character or word in images, on the other hand, has been among the prime objectives of computer vision for several decades. The novelty in this work is that it will take the 2D image of any format like jpeg, bmp, tiff, gif etc there is no specific format of the image. Therefore in this proposed algorithm the main three stages for HCR are preprocessing, feature extraction, classification**.**

HCR is divided into three stages:

1. Pre-processing 2. Feature Extraction 3. Classification

1. Pre processing

In the pre-processing stage we have to select one image as per our interest which is coloured one. We have to convert that image into gray scale image for better visualization of information stored in each pixel.

Create two masks one is for horizontal projection and other for vertical projection of gray scale image to detect & eliminate shirorekha of word.

2. Feature Extraction:

Now we have Segmented Image as our Binary Image. We have to crop each character from the Binary image word. Now we

have cropped characters and have to find feature extraction of each character using K-Means clustering (For our Database this is a best suitable method except Contour Extraction, region growing....etc.).

3. Classification

Classification of the Hindi characters is performed by the Euclidean distance method. Where for this type of algorithm we estimate the future vectors for the each and every train data and store it in a data base. The Euclidean method is a simple method but powerful enough for the detection of the characters. For the case of test data we estimate the vectors and then these vectors are separated from the train data.

OCR finds wide applications as a telecommunication aid for the deaf, postal address reading, direct processing of documents, foreign language recognition etc. This problem has been explored in depth for the Latin script.

Large intra-category variations of appearances and instantiations within character classes turn learning category models into a key challenge. Therefore, common characteristics of a character have to be captured while offering invariance with respect to vari-abilities or absence of these features. In order to determine the pattern K mean clustering algorithm is used. Clustering algorithm plays an important role in Hindi character identification. The currently best approaches to recognition of hindi character is K mean clustering algorithm.

II. Proposed algorithm for HCR

The main stages in proposed HCR system are

1. Preprocessing
2. Segmentation
3. Feature Extraction
4. Classification

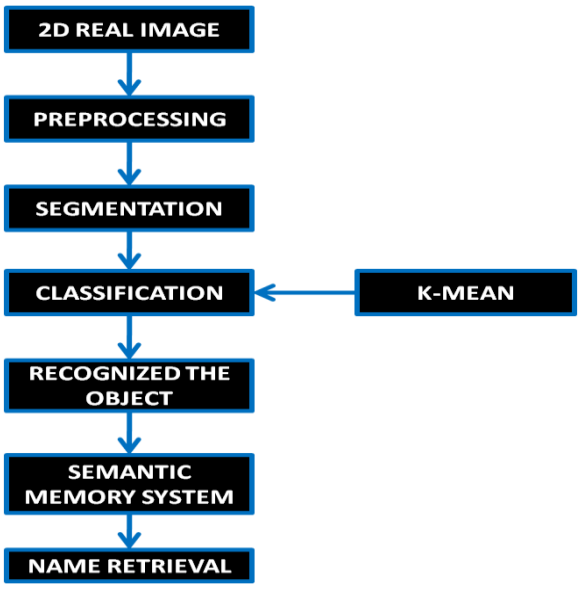

Figure1: Steps of proposed algorithm of HCR

STEP 1: Preprocessing

The proposed algorithm starts with taking the 2D image of any format like jpeg, tiff, gif etc. The preprocessing is done on that image. In preprocessing, from the input image, the discontinuity and the distortion are removed or we can say that the noise can be removed. It consists the following:

I. **Determination of the size of the input image:** determine the approximate dimensions of the characters by forming a fit rectangular boundary around the character.

II. **Distortion Removal :**We use thickening, thinning and pruning for removing distortions [1]. The image is thickened first and then thinned to convergence. This gives us a smooth one-pixel wide image [2] of the character, which is pruned to remove the small projections resulting from the thinning algorithm. Small characters should be distortion free.

Noise reduction can be done by many techniques like
a) Filtering
b) Morphological operations
c) Noise modeling

STEP 2: Segmentation

The very important step in HCR is segmentation. If we use significant segmentation techniques it will enhance the efficiency of HCR system. Segmentation is important because we can extract the basic unit of the script which is the Hindi character.

Segmentation is also necessary because there are many touching characters i.e. characters are touched with some other characters by the means of shirorekha, the system will not recognize the touching character therefore segmentation is important.

By doing horizontal scanning we can segment the horizontal line from the word. Then the individual characters are segmented. For separating characters first the headline will remove it is done by converting the black pixels to the white pixels. When the head line is removed the word is divided into three zones- upper zone, middle zone, and lower zone. Then the vertical scanning is done and the individual characters are separated from each zone.

STEP3: Feature Extraction

Feature extraction extract the feature or a data from the input image which is useful for classifier for recognize the character. Feature extraction involves representing a handwriting text by a set of discriminative features. The feature representation is based on removal of certain types of information from the image [3]. Feature extraction is necessary step for classification.

STEP4: Classification

Classification is based on the k-means clustering algorithm. Classification is used for making decisions. The classifier's performance depends on the feature that is extracted.

K-means clusteringalgorithm - is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. This procedure gives a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters). The motive is to define k no. of centroids, one centroid for each cluster. Because different locations causes different results that's why these k centroids should be placed in a cunning way. So, the better choice is to place them as much as possible far away from each other. The next step which is followed is to take each point which is belonging to a given data set and associate it to the nearest centroid. When no point is remaining, the first step is completed. Now there is a need to re-calculate k new centroids as bary centers of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. Finally, this algorithm aims at minimizing an objective function, in this case a squared error function.

STEP5 : Recognize object and stored in the memory

By performing the previous steps the object is recognized and i.e. the Hindi character. The recognized character then stored in the database and can be used further or recognizing the other characters which are similar to it. This is the final step of the algorithm.

III. Conclusion

The proposed algorithm for HCR using k-means clustering algorithm is described. By using this proposed algorithm we can recognize the Hindi characters which is very necessary in today's world. The shirorekha detection and removing it, noise reduction, segmentation for separating individual characters, feature extraction and the classification by using k-means algorithm are the main steps of this algorithm. Hindi is India's national language, all the government work are done in Hindi language. Paper work will not convenient for today's generation, all official works will done digitally. Therefore HCR system is very important. We can digitally save our paper work by scanning them because paper will not remain lastly.

REFERENCES

[1] E.R. Davies and A.P. Plummer, "Thinning Algorithms: A critique and new Methodology" ,Pattern Recognition14, [1981]: 53-63

[2] S. Arora, D.Bhattacharya, M. Nasipuri, L.Malik, "A Novel Approach for Handwritten Devanagari CharacterRecognition" in IEEE –International Conference on Signal And Image Processing, Hubli, Karnataka, Dec 7-9, 2006

[3] VedguptSaraf, D.S. Rao, "Devnagari Script Character Recognition Using Genetic Algorithm for Get Better Efficiency" in International Journal of Soft Computing and Engineering (IJSCE), April 2013

[4]Rahul KALA1, Harsh VAZIRANI2, Anupam SHUKLA3 and Ritu TIWARI4, "Offline Handwriting Recognition using Genetic Algorithm",
IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 2, No 1, March 2010.

[5]Shabana Mehfuz1, Gauri Katiyar2, "Intelligent Systems for Off-Line Handwritten Character Recognition: A Review ", International Journal of Emerging Technology and Advanced Engineering, Volume 2, Issue 4, April 2012

Author Profile

Meha Mathur received the B.Tech degrees in Information Technology from Vyas Institute of Engineering and Technology in 2012 and pursuing M.Tech degree from Amity University Rajasthan. Now working on the implementation of HCR.

Conflict of interest statement
I declare there is no conflict of interest.