7th National Conference on
"Role of Engineers in Nation Building" organized by VIVA Institute of Technology, Mumbai
(1st and 2nd March 2019)
International Journal of Engineering Research & Science (IJOER)
ISBN: [978-93-5321-411-1] ISSN: [2395-6992] [Vol-5, Issue-3, March- 2019]

# A Survey on Hindi Character Recognition

Sahil Angwalkar[1], Suhas Chaubey[2], Raj Mehta[3], Pallavi Vartak[4]
Department of Computer Engineering, Mumbai University, INDIA

***Abstract*—***The Devanagari Script is used in many Indian languages. Hindi which is the most commonly used language in India also uses the Devanagari script. Recognition of Hindi Characters is an important area of research because of various applications like Library automation, publication house, manuscripts, Granths and various other documents. A lot Machine Learning and Deep Learning approaches have been used for recognition of Hindi as well as English characters and digits. This paper presents a survey of various approaches used for Hindi Character Recognition.*

***Keywords—Character Recognition, Devanagari, Deep Learning, Machine Learning, Neural Network.***

## I. INTRODUCTION

The heritage of India in literature and languages is richest in the world. India is a language wise very much diverse. We have 22 official languages in India of which English and Hindi is most widely used. So making such system which can recognize these languages is important [4].

When computers became popular in the 1950s, huge amount of data was manually processed into the computers. In various organizations like finance, insurance, government and public utilities, millions of documents need to be processed and thus manual processing becomes a cumbersome task. All this necessities acknowledged the need for Character Recognition [5].

A lot of work is done for English Character Recognition but the same cannot be said about Hindi Characters. Hindi language is written in Devanagari script which is used by various other languages. The Devanagari script is widely used for documentation, Literature, etc. in India. So recognition of Hindi Characters is an important area of research. Hindi language has 14 modifiers, 13 vowels and 34 consonants. The vowels are called 'Swar' and consonants are called 'Vyanjan' in Hindi language [5]. There are basically three major steps in Character Recognition: Pre-processing, Feature Extraction and Classification. In the pre-processing step, the noises and distortions in the images are removed. The features on the basis of which the characters will be recognized are extracted in the Feature Extraction step. Various classification algorithms are then used to recognize the images [1].

## II. LITERATURE SURVEY

### 2.1 Hindi Handwritten Character Recognition using Multiple Classifiers [1].

After gaining knowledge of languages, humans can easily recognize handwritten characters. This knowledge needs to be transferred to machines so that it can automatically recognize characters. In this paper, there are four phases for Character Recognition viz. Database Creation, Preprocessing, Feature Extraction and Classification. Database of 4,428 characters is created by taking 108 samples of each character. In Pre-processing, aspect ratio adaptive normalization is used for normalizing the characters and Median Filter is used to remove noise from input characters. In Feature Extraction, two features are used namely Histogram of Oriented Gradients and Projection profile histogram. These two features are then fed to multiple classifiers like Quadratic SVM, k-NN, weighted k-NN and Bagged Trees. Quadratic SVM gave the best results.

### 2.2 Hindi Character Recognition using RBF Neural Network and Directional group feature extraction technique [2].

In this proposed system, Radial Basis Function (RBF) Neural Network is used for Hindi Character Network. Initially, the input image of the character is Binarized. Then gradient feature extraction is performed using Sobel Operator. The different range of gradients is mapped into different directional group values (0-8). The Radial Basis Function of one input, one hidden and one output layer is used for training the Neural Network. To train the Radial Basis Function (RBF) Network, gradient descent training algorithm has been used. The results obtained using RBF Neural Network is then compared with results of Back Propagation Neural Network. Comparative results show that RBF Neural Network gives slightly less recognition accuracy but it reduces the training and classification time.

International Journal of Engineering Research & Science (IJOER)
ISBN: [978-93-5321-411-1] ISSN: [2395-6992] [Vol-5, Issue-3, March- 2019]

### 2.3 Recognition of Offline Handwritten Hindi Text using Middle Zone of words [3].

In the proposed system, a segmentation based approach is used for is used to recognize Offline Handwritten Hindi Text. The text was scanned at 300 dpi and reduced to 35% size in paint for faster execution. No preprocessing is done in this proposed system. Vertical Projection method is used to segment the text into words. Then the upper and lower modifiers are segmented from words. Finally, consonants, half characters and joint words are segmented. After each character is segmented, it is given a unique number which is used as a feature during recognition phase. These features are then fed to a classifier. The main aim of this paper is to reduce training time using segmenting method.

### 2.4 Handwritten Hindi Numeric Character Recognition and comparison of Algorithms [4].

The heritage of India in Literature and Languages is the richest in the world. Hindi and English are the two most important and popular languages of India. Thus recognition of the characters of these languages is a great field of research. In this paper, a system for Hindi Numeric Character Recognition is proposed. Database is constructed by taking handwritten data from over 100 writers. For feature extraction two algorithms are used namely Histogram of Oriented Gradients (HOG) and Principal Component Analysis (PCA). PCA algorithm finds the features with maximum variance and these features are fed to the classifier. The classifiers used in this system are: K –Nearest Neighbors (KNN) and Support Vector Machine (SVM). Different combinations of Feature Extraction algorithms and Classifiers are used and compared. The combination of HOG and KNN give best results.

### 2.5 Handwritten Hindi Character Recognition using K-means clustering and SVM [5].

The Devanagari script is used in many Indian languages. Hindi language also comes under Devanagari script. In the proposed system, the Hindi Character Recognition is done by using a three step procedure. First step is preprocessing. In this step, the image is first converted into Binary image which gives two outputs viz. vertical and horizontal binary image. These two outputs are combined by And operation and the horizontal line is removed. The next step is Feature Extraction. In this step, K-means clustering is used. The image is divided into clusters and pixels under each cluster are combined together to form a feature vector. This vector is then fed to the third step which is classification step. Support Vector Machine (SVM) is used as a classifier in this system.

### 2.6 A Survey of Offline Handwritten Hindi Character Recognition [6].

The character recognition problem can be classified in to two categories viz. Printed character recognition and handwritten character recognition. The handwritten character recognition problem is further classified into two categories: Offline Handwritten character recognition and online character recognition. In this paper, an overview of what methods have been used so far for Hindi character recognition is provided. The methods used so far for feature extraction are: Structural features, Binary Vector of Image, Clonal Selection algorithm, etc. The classifiers used are Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Back Propagation Neural Network (BPNN), etc.

### 2.7 Optimized and Efficient Feature Extraction Method for Devanagari Handwritten Character Recognition [7].

The three main tasks in handwritten character recognition are Image segmentation, Feature Extraction and classification. Feature extraction is a very important step in Character recognition. In this paper, an optimized and efficient feature extraction method is proposed. Initially in the Preprocessing step, the Skeletonization and Universe of discourse are used to reduce feature Extraction time. Then two different features are extracted namely: Regional and Gradient Features. The regional features are the number of holes and number of objects in input image and the ratio of white pixels in the image to total pixels. Gradient features are the directional change in the intensity of pixels. These two features are combined together and a feature vector is constructed. Better accuracy is achieved using this feature vector.

### 2.8 Optical Character Recognition (OCR) Systen for Roman Script & English language using Artificial Neural Network (ANN) Classifier [8].

Recognizing character from scanned image is a very difficult task. But we need all the data to be in the digital format for the purpose of record keeping. In this paper, a new approach is proposed using the concept of Artificial Neural Network (ANN) and

International Journal of Engineering Research & Science (IJOER)
ISBN: [978-93-5321-411-1] ISSN: [2395-6992] [Vol-5, Issue-3, March- 2019]

Nearest Neighbors. Three layers used in the system are: Input layer, Hidden layer and Output layer. Input layers consists of input given from segmented characters, Hidden layer consists of trained neurons and Output layer consists of output neurons which provides recognized output. Initially methods such as quantization and resizing is used in the preprocessing process. Then canny edge detector filter is used to detect where exactly the text is in the image. Nearest Neighbor classification approach is used to recognize the characters.

### 2.9 Deep Random Vector Functional Link Network Handwritten Character Recognition [9].

The Artificial Neural Networks have a history of several decades but it did not have the computing power which the computers of current generation have. Therefore some of this Networks are rebranded and represented again. In the proposed system, two approaches are tested on four different databases. The two approaches are Random Vector Functional Link Network (RVFL) and Extreme Machine Learning (EML). Random Vector Functional Link Network (RVFL) is an artificial feed forward neural network with single hidden layer for both classification and regression. Extreme Machine Learning (EML) consists of different variations of Random Vector Functional Link Network. Extreme Machine Learning (EML) uses mean and median of image pixels as a feature vector to be fed to multi-classifier system of Random Vector Functional Link Network (RVFL)/ Extreme Machine Learning (EML).

### 2.10 Recognition of Cursive English Handwritten Characters [10].

In this paper, an approach for cursive English handwritten character recognition is proposed using Convex Hull Algorithm and Support Vector Machine (SVM) for classification and recognition purpose. Initially the image is acquired by scanning the document. Then noise removal, skew correction, cropping and resizing, normalization, thinning, binarization, skeletonization techniques are used for preprocessing the input image. The document is then segmented first into lines then words and lastly characters. Then Convex Hull algorithm is then used for feature extraction and these features are fed to the Support Vector Machine (SVM) for classification and recognition purpose.

### 2.11 Optical Character Recognition using KNN on custom image dataset [11].

In this paper, an efficient method for training classifier using custom image is proposed. The Optical Character Recognition (OCR) system extracts features from the input image to recognize character specifically alphabets and numbers. Initially, the image is converted into grayscale, blurred, threshold and flattened image so that it is easier for the system to understand features. These features are then stored in a numpy array. This array and the labeled array is stored in a text file. This text file is used for training the K-Nearest Neighbors (KNN) classifier. Using K-Nearest Neighbors (KNN) the character with the nearest features is recognized. A hidden layer is used to extract features and the classifier is trained on the basis of it.

### 2.12 Optical Character Recognition using Back Propagation Neural Network [12].

In the proposed system, Artificial Neural Network (ANN) using feed forward Neural Network is used for English character recognition. One of the important factors that degrade performance of recognition system is Noise which is being considered in this paper. The proposed is divided basically into two sections: training section and recognition section. In the training section, first the image is acquired and preprocessing techniques like binarization and Thresholding are performed on it. The columnized matrix is used for feature extraction in this paper. The Artificial Neural Network (ANN) is trained on the basis of this feature. This trained network is used for character recognition.

## III. ANALYSIS TABLE

The Analysis table is the summary of the studied research papers and the different techniques used on Sentiment analysis of code-mixed text.

International Journal of Engineering Research & Science (IJOER)
ISBN: [978-93-5321-411-1] ISSN: [2395-6992] [Vol-5, Issue-3, March- 2019]

**TABLE**

| Sr. No. | Title of Paper | Technique/ Methods | Accuracy |
|---|---|---|---|
| 1. | Hindi Handwritten Character Recognition using Multiple Classifiers [1] | HOG and Projection profile histogram for feature extraction and Quadratic SVM as classifier. | 96.6% Accuracy |
| 2. | Hindi Character Recognition using RBF Neural Network and Directional Group Feature Extraction Technique [2] | Sobel Operator is used for feature extraction and Gradient Descent algorithm is used to train RBF neural network. | 92% Accuracy |
| 3. | Recognition of Offline Handwritten Hindi Text using Middle Zone of words [3] | Segmentation is performed using strip wise Vertical projection method. | 87.35% Accuracy |
| 4. | Handwritten Hindi Numeric Character Recognition and comparison of algorithms [4] | HOG and PCA for feature extraction and SVM and KNN are used as classifiers. | 95.27% Accuracy |
| 5. | Hand written Hindi Character Recognition using K-means clustering and SVM [5] | K-means clustering is used for feature extraction and SVM is used for classification. | 81.7% Accuracy |
| 6. | A Survey of offline Hindi Handwritten Character Recognition [6] | Structural Features and SVM, KNN and BNN are used for classification. | 96.14% Accuracy |
| 7. | Optimized and Efficient feature extraction method for Devanagari handwritten character recognition [7] | Structural features, Regional features, Gradient features, etc. are combined together. | 94% Accuracy |
| 8. | OCR system for Roman Script and English language using ANN classifier [8] | Artificial Neural Network (ANN) is used. | 98.89% Accuracy |
| 9. | Deep Random Vector Functional Link Network for Handwritten Character Recognition [9] | Random Vector Functional Link Network (RVFL)/ Extreme Machine Learning (EML) | 98.64% Accuracy |
| 10. | Recognition of Cursive English handwritten characters [10] | Convex Hull algorithm is used for feature extraction and SVM is used for classification. | Not mentioned |
| 11. | Optical Character Recognition using KNN on Custom Image Dataset [11] | KNN algorithm | Not mentioned |
| 12. | Optical Character Recognition using Back Propagation Neural Network [12] | Back Propagation Neural Network (BPNN) | 96% Accuracy |

## IV. CONCLUSION

The Devanagari script is used in many Indian languages. Hindi also come under the Devanagari script and is avery popular language in India, thus recognition of Hindi Characters is an important area of research. Various Machine Learning and Deep Learning techniques such as K-means Clustering, Support Vector Machine (SVM), RBF Neural Network, Artificial Neural

International Journal of Engineering Research & Science (IJOER)
ISBN: [978-93-5321-411-1] ISSN: [2395-6992] [Vol-5, Issue-3, March- 2019]

Network (ANN), K-Nearest Neighbours (KNN) and Back Propagation Neural Network have been used for Character Recognition. In this paper, different machine learning and deep learning techniques have been studied and analysed. It is evident from the studied papers that deep learning approaches have given better accuracy. In future, using Convolutional Neural Network (CNN) may help in further increasing the accuracy.

## REFERENCES

[1] Madhuri Yadav and Dr.Ravindra Purwar,"Hindi Handwritten Character Recognition using Multiple Classifiers", 2017 7th International Conference on Cloud Computing,Data Science & Engineering, pp. 149-154.

[2] Dayashankar Singh, Dr.J.P.Saini, Prof. D.S Chauhan,"Hindi Character Recognition using RBF Neural Network and Directional Group Feature Extraction technique", IEEE 2015, pp.

[3] Naresh Kumar Garg, Lakhwinder Kaur, Manish Jindal,"Recognition of Offline Handwritten Hindi Text using Middle Zone of the words", 2015 IEEE ICIS 2015, June 28-July 1 2015, Las Vegas,USA.

[4] Apoorva Choudhary and Mr. Roshan lalchokker, "Handwritten Hindi Numeric Character Recognition and comparison of algorithms", IEEE 2017, pp. 13-16.

[5] Ajay Indian and Karamjit Bhatia,"A Survey of Offline Handwritten Hindi Character Recognition", IEEE 2017, pp. 65-70.

[6] Akanskha Gaur and Sunita Yadav,"Handwritten Hindi Character Recognition using K-Means Clustering and SVM", 2015 4th International Symposium on Emerging Trends and Technologies in Libraries and Information Services.

[7] Mrs. Saniya Ansari and Dr.Udaysingh Sutar,"Optimized and Efficient Feature Extraction Method for Devanagari Handwritten Character Recognition", 2015 International Conference on Information Processing (ICIP) Vishwakarma Institute of Technology. Dec 16-19, 2015, pp. 11-15.

[8] Honey Mehta, Sanjay Singla, Aarti Mahajan," Optical Character Recognition (OCR) System for Roman Script & English Language using Artificial Neural Network (ANN) Classifier", International Conference on Research Advances, 2016.

[9] Hubert Cecotti,"Deep Random Vector Functional Link Network for Handwritten Character Recognition", IEEE 2016, pp. 3628-3633.

[10] Pritam Dhande and Reena Kharat,"Recognition Of Cursive English Handwritten Character", International Conference on Trends in Electronics and Informatics ICEI 2017, pp. 199-203.

[11] Tapan Kumar Hazara, Dhirendra Pratap Singh, Nikunj Daga," Optical Character Recognition using KNN on Custom Image Dataset", IEEE 2017, pp. 100-114.

[12] Shyla Afroge, Boshir Ahmed, Firoz Mahmud," Optical Character Recognition using Back Propagation Neural Network", 2nd International Conference on Electrical, Computer and Telecommunication Engineering (ICECTE) 2016.