Impact Factor (SJIF): 4.542

International Journal of Advance Research in Engineering, Science & Technology

e-ISSN: 2393-9877, p-ISSN: 2394-2444

Volume 4, Issue 4, April-2017

OPTICAL CHARACTER RECOGNITION FOR HINDI CHARACTER USING SVM

Hiren R. Vasava(PG Student)

Department of Information Technology, SSEC, Bhavnagar

Abstract* —Optical Character Recognition is an evergreen area of research and widely used in different Real Time applications. This paper proposes a system for recognizing offline Handwritten Hindi Characters using Support Vector Machine(SVM). Data-Set are collected from different writers Preprocessing is done using various morphological operation and thinning technique. All the Training was done using self-created database consist of 50 samples per character from 50 different individuals. For Feature Extraction View Based Technique is used. Support vector machines found to be very efficient and robust in handling large amount of features, SVM was used for the purpose of classification

Keywords- Optical Character Recognition(OCR), Feature Extraction, Support Vector Machine(SVM), Classification, Hindi Character recognition ,Preprocessing

I. INTRODUCTION

OCR, which is an abbreviation of Optical Character Recognition, converts the scanned image into usable format. These scanned images can be printed character images, handwritten character images. Mainly it is used at the time of data entry from data source, which is written on paper. OCR is mainly categorized into two parts, first one is, Online character recognition and the second one is Offline character recognition. In Online character recognition, characters are recognized at the time of writing and it uses the time stamp process for this. Offline character recognition uses the image of characters and converts them into computer understandable format. Offline character recognition can be done on both printed and handwritten texts. Handwritten character recognition is more difficult in comparison to printed character recognition because of diversity in.

In this paper, handwritten Hindi character recognition is presented. Hindi comes under Devanagari script. Hindi is India’s national language and is very popular. There are 14 modifiers(Matras) and 13 vowels, shown in Fig 1(a) and 34 consonants, which are shown in Fig1 (b), in Hindi language. Hindi vowels are called ‘Swar’ in Hindi language and Hindi consonants are also called ‘Vyanjan’ in Hindi language.

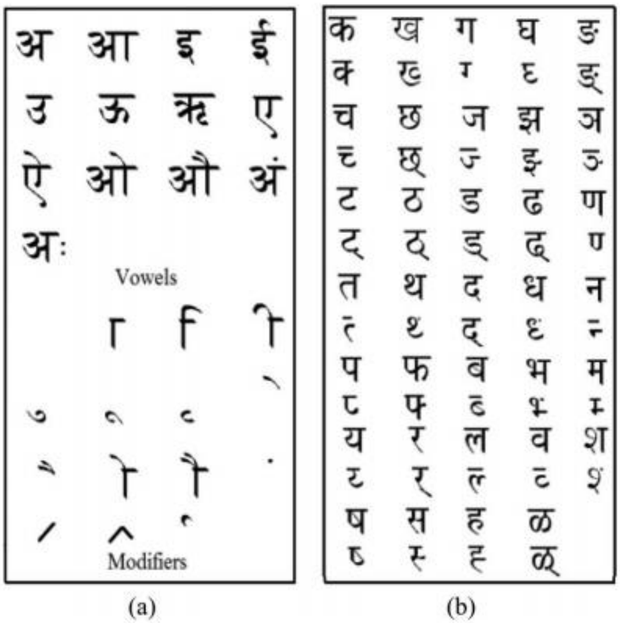


Figure 1 (a) Vowels and modifiers of Hindi Language. (b) Consonants and their corresponding half forms (shown below the consonants) in Hindi Language.

For character recognition researchers use many techniques. In character recognition, first the preprocessing of scanned image is required to remove noises from scanned image. Raw image can have different type of noises, distortion etc. Removal of noises from scanned image makes the recognition of characters easy. After preprocessing of the image, the special quality of character is extracted. This process is called feature extraction process and the special quality is called feature. Many feature extraction techniques are used by various researchers, like structure features, contour features, ring features, Zernike features, ink based features, gradient features, global features etc. Structural details like endpoints, intersection of line segments, loops, curvatures, segment lengths, etc. are also used which describes the geometry of the pattern structure as feature.

Image is divided into segments and structure features are extracted from each segment individually .On the basis of features, classification process is executed. Classification is the process in which objects are differentiated and categorized into classes. For classification process, different types of techniques are used by various authors, like, Neural Network, Fuzzy Logic, HMM, Support vector machine and hybrid techniques too. KNN-SVM, which is the hybrid approach of K-Nearest Neighbor and Support Vector Machine, gives the results in the specialization of SVMs in the local areas around the separation surface. Fuzzy logic is also used for classification purpose. Another hybrid approach is used as a classifier which combines SVM & MQDF

II. Related Work

Many researchers have done work on OCR and contributed in field of both online and offline, but very less work would be found on HCR especially for Hindi Language.
Akansha Gaur[1] took k-Means clustering as a feature and used Euclidean distance and support vector machine as a classifier to recognize Hindi characters and obtained 95.86% accuracy with svm. Ashok Kumar Bathla[2] focuses on Hindi text and its challenges in character segmentation because of broken character, overlapping characters, touching characters. Nisha Goyal[3]focuses on the feature extraction technique. Feature extraction in any handwritten script is very important part of optical character recognition. Zoning is used for the feature extraction technique to recognize handwritten devanagari script. Binny Thakral[4]focuses on the Segmentation part. Segmentation is the indispensable and most difficult part of the optical character recognize process.It gets to be additionally difficult with handwritten text due to varieties in writing style and presence of abnormalities. In this paper new techniques is shown for the segmentation of conjuncts, and overlapping characters in Hindi language. Arjun Singh[5] focuses to recognizing offline Handwritten Devanagari Characters using artificial neural network and support vector machine as classifiers and the results are compared Various Feature Extraction Technique is used like chain code, zone based Centroid , background directional distribution and distance profile features are applied to the pre processed images.

Shilpy Bansal[6]focuses on the offline Handwritten Gurmukhi character recognition. Neighborhood foreground pixels density technique is used for feature extraction technique. Some insignificant feature values so to reject those we have used a dimensionality reduction technique namely Principal component analysis(pca).Maximum 91.95% accuracy is achieved with SVM classifier by using 10 fold cross validation test method. Ashutosh Aggarwal[7]focuses on the offline handwritten Gurmukhi characters. In these paper two sets of features based on gradient and curvature of character image are computed and extracted features are then used together to form a composite feature vector containing both gradient and curvature information. Support vector machine is used for classification purpose. R.Ramanathan[8]focuses on new technique of optical character recognition using Gabor filter and support vector machines. The model proposed is trained and validate two languages- Tamil and English. Shailendra kumar Shrivastava[9]in this paper Support vector machine is used for classification in optical character recognition for Devanagari Numeral. .Database is constructed by implementing Automated Numeral Extraction and Segmentation Program(ANESP).Linear kernel function is used in svm which give 99.48% overall recognition rate. N. Shanthi[10]This paper describes a system for recognizing offline handwritten Tamil characters using support vector machine. Data are collected from different writers. The svm is tested for the first time to recognize handwritten Tamil characters and give 82.04% accuracy.

III. Proposed Methodology

The flow diagram of proposed method is shown in Fig 2, which is divided in to main 3 parts: Pre-processing Phase, Feature Extraction Phase and Classification Phase

3.1 Create Image

First of all Image is created by scanning the handwritten document and then it is converted into required format like jpeg or png etc.

3.2 Pre-processing

In preprocessing part scanned image is converted to binary and various other techniques to remove noise and to make ready for the feature extraction. This part includes segmentation, normalization, skeletonization and filtration which types of technique will best suit is depended on by mechanism we used. Some Preprocessing techniques are listed below.

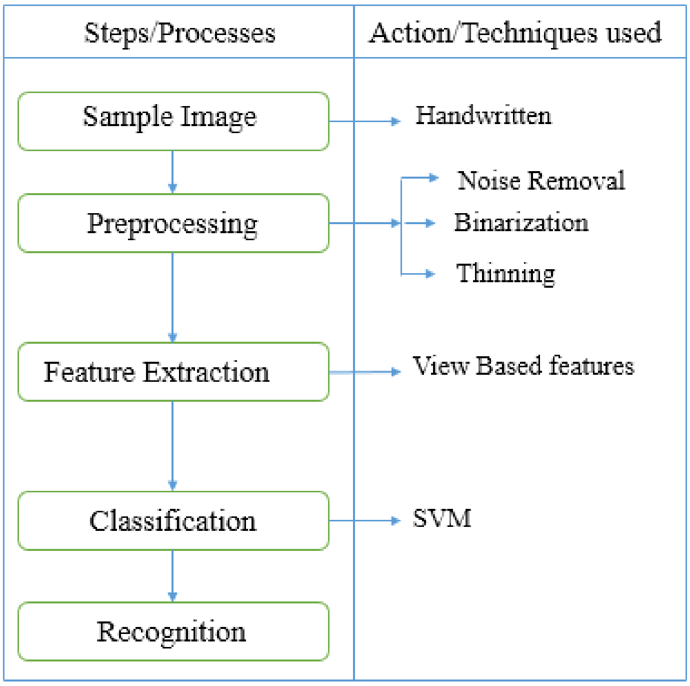


Figure 2: Proposed OCR System Architecture

3.2.1 Gray Scale

If the input image is colored then it may be required to first convert in the gray scale, later convert to binary image.

3.2.2 Binarization

Binarization converts colored or gray scale image into binary image. The image scanned and pre-processed is binarized, converted in the 0 and 1 form. The place where part of image is not shown or it is white is displayed as 0 and the part where image is shown or it is black is displayed as 1.

3.2.3 Noise removal

For noise removal, smoothing operation is used. It is used to blur the image and reduce the noise or to straighten the edge of the character.

3.2.4 Thinning

Thinning is an important step in optical character recognition. Main purpose of this step is to delete redundant information and to retain the characteristic feature same. It is applied to find skeleton of character.

3.3 Feature Extraction Technique

This is the very important step of optical character recognition process. If right feature extraction method is used, then it will make easy to recognition of individual character. Some attribute extraction technique are Template matching, Contour Profile, Projection Histogram, Moments calculation, Zoning and Deformable Templates. In this work we are using View based feature extraction technique is used.

3.3.1 View based feature

This method is based on the fact that for correct character recognition a human usually needs only some information about it, its shape and contour. This feature extraction method, which works on scaled, thinned binarized image, examines four “views” of each character extracting from them a characteristic vector, which describes the given character as shown in fig. 3. The view is a set of points that plot one of four projections of the object (left, right, top, bottom) it consist of pixels belonging to the contour of the character and having extreme values of each block. Here for 5×5 blocks we get 5×5×8 = 200 features for recognition 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. In the considered examples, eleven uniformly distributed characteristic points are taken for each view. 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 44 obtained values the feature vector is created to describe the given character, and which is the base for further analysis and classification.

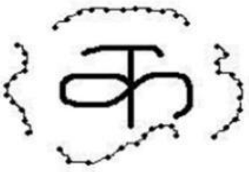


Figure 3: Selecting characteristic points for four views

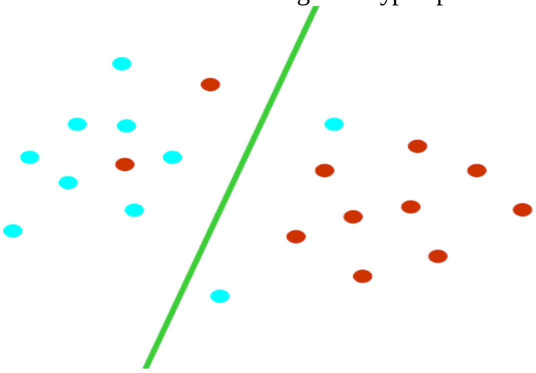
3.4 Classification Technique

Classification is the last step of Optical Character Recognition. Based on feature extraction technique classification is performed as to which class the character belongs. Support vector machine is used as a classification technique in this work which classifies the class of the character. Some classification Technique used in formerly developed OCR is Support Vector Machine, K-Nearest Neighbors, Neural Network, Decision Tree classification and Bayesian classification.

3.4.1 Support Vector Machine

For classification of characters Support Vector Machine (SVM) is used. SVM is based on the supervised learning that used for analyzing of data. SVM uses hyper-plane for classification. Hyper-plane with the maximum margin of separation of hyper-plane and closest data point, is used as a decision surface.

This Optimal Hyper-plane gives the output. Different types of kernels are used in SVM: Linear, RBF, Quadratic, Polynomial and MLP. Here RBF kernel is used for classification with SVM. In Fig 4, this green line is the optimal hyper-plane which is separating two sets with maximum margin of hyper-plane and nearest data points.



: Figure 4: SVM with optimal separating hyper-plane

IV. RESULTS

In this research work, data is written by 50 people from various fields such as Education, Service, Industry etc..of different ages ranges from 10 years boy to 60 years gentleman. For Implementation MATLAB is used as a tool. For feature extraction View based technique is used and SVM is used as classification Method. Some Characters are continuously giving good performance and some are bad. Using this technique, the recognition result is achieved 97.55%.

$$\% \text{ of performance} = \frac{\text{Total no. of Recognized characters}}{\text{Total characters for testing}}$$

V. CONCLUSION

From the result it can be concluded that combination of svm Classifier and View based feature extraction approach is best method for the recognition of offline handwritten characters. The future work may involve the recognition of complete sentences as well as speech can be synthesized for the individual character that is recognized by the system.

REFERENCES

[1] Akansha Gaur, Sunita Yadav ”Handwritten Hindi Character Recognition using K-Means Clustering and SVM”, 4th International Symposium on Emerging Trends and Technologies in Libraries and Information Services IEEE- 2015

[2] Ashok Kumar Bathla, Sunil Kumar Gupta, Manish Kumar Jindal “Challenges in recognition of Devanagari Scripts due to Segmentation of Handwritten text” International Conference on Computing For Sustainable Global Development(INDIACom)IEEE-2016

[3] Nisha Goyal, Er. Shilpa Jain “Optimized Hindi Script Recognition using OCR Feature Extraction Technique”,International Journal of Advanced Research in Computer and Communication Engineering. vol. 4, issue 8, 2015

[4] Binny Thakral, Manoj Kumar “Devanagari Handwritten Text Segmentation for Overlapping and Conjunct Characters-A Proficient Technique”, IEEE-2014

[5] Arjun Singh, Kansham Angphun Maring ”Handwritten Devanagari Character Recognition using SVM and ANN”, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, issue 8, 2015

[6] Shilpy Bansal, Mamta Garg, Munish Kumar “A technique for Offline Handwritten Charcter Recognition”. International Journal of Computing and Technology, Volme 1, Issue 2, 2014

[7] Ashutosh Aggarwal, Karamjeet Singh “Handwritten Gurmukhi Charcter Recognition”, IEEE International Conference on Computer, Communication and Control(IC4-2015)

[8] R.Ramanathan, S.Ponmathavan, N.Valliappan, Dr. K.P.Sonam ”Optical Character Recognition for English and Tamil Using Support Vector Machines”, International Conference on Advances in Computing, Control, and Telecommunication Technologies, IEEE-2009

[9] Shailedra Kumar Shrivastava, Sanjay S. Gharde “Support Vector Machine for Handwritten Devanagari Numeral Recognition ”, International Journal of Computer Applications(0975-8887), vol. 7- No. 11, 2010

[10] N. Shanthi, K. Duraiswamy “A novel SVM-based handwritten Tamil character recognition system”, Springer-2009