Offline classifier for Handwritten Devanagari script recognition

KAPIL BAMNE, department of communication Eng. ,uecu, Ujjain, indore,india,
Dr.NEHA SHARMA,department of communication Eng.,uecu,Ujjain,Ujjain,India

Abstract— **Hindi is the most popular language in India and the world's third most commonly used language after Chinese and English. Hindi script presents great challenges to design due to the large number of letters present in the script and the sophisticated way, in which they combine. Automatic recognition of handwritten characters has long been a goal of many research efforts in the pattern recognition field. Written text image may be sensed "off line" from a piece of paper by optical scanning (optical character recognition). Devanagari script has 14 vowels and 33 consonants. Vowels occur either in isolation or in combination with consonants. The net result is that there are several thousand different shapes or patterns, which makes Devanagari OCR more difficult to develop. Here focus is on the recognition of offline handwritten Hindi characters that can be used in common applications like commercial forms, bill processing systems ,bank cheques, passport readers, offline document recognition generated by the expanding technological society . A comparative study of various Handwritten recognition methods is being taken into account .**

Index Terms— **Classifier, feature extraction, OCR, Segmentation.**

I. INTRODUCTION

A. Overview of Devangiri Script

There are 3 characteristics [1], [2], [3] of the Devanagari script which distinguish it from most of the other scripts.

Shirorekha: The Shirorekha is a horizontal line which is drawn at the top of each character and extends throughout the word in Devanagari. Though, for few specific characters the shirorekha is written only partially on top.

Modifiers: Modifiers are strokes that are attached to the basic characters in order to change the pronunciation of the

Character: A basic character with a modifier becomes a modified character, which is then concatenated with other basic/modified characters to form a word.

Conjunct Character: Here a character occurs in its half form and is attached to another basic character, which is complete. This combination is called as conjunct character.

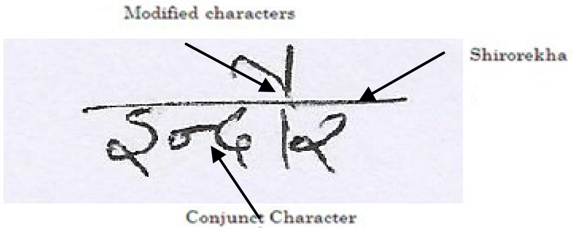


Fig 1. Characteristics of Devanagari Character

B. Data Acquisition:

Data acquisition is the first phase in any image processing or pattern recognition task. We have designed special kinds of forms to collect the handwritten samples from the writers. The form contains different boxes in which writer has to write all the name of the cities and compound characters words in the lexicon in a specified order. Most of the writers were from the age group 16 to 25. There was no restriction imposed on the writer regarding the style and speed of writing.

C. Digitization

Digitization is a process to convert printed documents into digital images. It is done by scanning the documents using a scanner. To scan the documents Deskjet Scanner is used with a resolution of 300 dots per inch (dpi). The scanner converts the hard-copy of document into gray-scale images.

1. Preprocessing

Preprocessing is an important step of applying a number of procedures for smoothing, enhancing, filtering, etc. for making a digital image usable by subsequent algorithm in order to improve their readability for optical character recognition software. The system performs character recognition by exploring the feature of character matching for its ability to recognize handwritten Devanagari Script.

i) A database of Hindi handwritten character is created in different handwritings from different peoples.
ii) Preprocessing of training image.

2. Binarization

Since the information contained in the text image is bi-level (text and background), we convert the gray scale images into binary images, and this process is called binarization. Binarization should be done carefully; otherwise it may lead to breaks in characters.

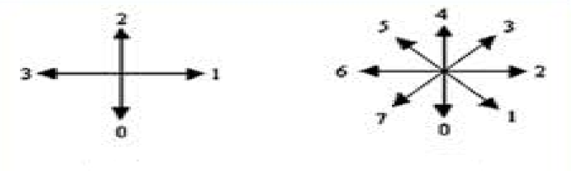
3.Segmentation

Character segmentation [4], is an operation that seeks to decompose an image of a sequence of characters into sub images of individual symbols. It is one of the decision processes in a system for character recognition. Its decision, that a pattern isolated from the image is that of a character (or some other identifiable unit), can be right or wrong. It is wrong sufficiently often to make a major contribution to the error rate of the system.

It is one the most important process, [5]that decides the success of character recognition technique. It is used to decompose an image of a sequence of characters into sub images of individual symbols by segmenting lines and words.

4. Feature Extraction

Feature extraction and selection can be defined as extracting the most representative information from the raw data, which minimizes the within class pattern variability while enhancing the between class pattern variability. For this purpose, a set of features are extracted for each class that helps distinguish it from other classes, while remaining invariant to characteristic differences within the class. Chain codes are used to represent the boundary of an object composed of pixels of regular cells by connected sequence of straight-line segments of specified length and direction.



.

4 directional 8 directional

Fig 2 chain codes

5. Classifier

The feature obtained from previous phase is assigned a class label and recognized using supervised and unsupervised method. The data set is divided into training set and test set for each character. Character classifier can be one or more of the following, Bayes classifier, Nearest neighbor classifier, Radial basis function, Support vector machine, MLP, Quadratic classifier, Linear, Modified discriminant functions, Gaussian distribution function, KNN, and Neural networks with or without back propagation. A number of classification methods were purposed by different researchers some of these are template matching, statistical methods, syntactic methods, artificial neural networks, kernel methods.

Classification is performed based on the extracted features. Handwritten Character Recognition systems extensively use the methodologies of pattern recognition, which assigns an unknown sample to a predefined class. Numerous techniques for PCR are investigated by the researchers.OCR classification techniques can be classified as follows:

1.Template Matching
2.Statistical Techniques
3.Neural Networks
4.Support vector Machine (SVM)algorithms
5.Combination classifiers

In Fuzzy logic and Neural Networks, we have to adjust weights and Number of hidden layers in order to achieve sufficient recognition rate. So, Neural network and Fuzzy logic are not capable to get high recognition rate. So, we are using Neuro-fuzzy integrated system to achieve high recognition rate.

II COMPARATIVE STUDY

This section describe the comparative study of various research work presented up till now. Recognition of handwritten characters has been a popular research area since 1970. The literature survey [6] carried out related to technology impact in the study of different text recognition techniques use on different languages of printed and handwritten scripts. Research in Indian offline character recognition started with the recognition of printed characters, irrespective of the script and then extended to the recognition of handwritten numbers and characters in many Indian scripts including Devanagari.

In a novel based approach [7], feature vector is constituted by accumulated directional gradient changes in different segments, number of intersections points for the character, type of spine present and type of shirorekha present in the character. One Multi-layer Perceptron with conjugate-gradient training is used to classify these feature vectors. This method is applied to a database with 1000 sample characters and the recognition rate obtained is 88.12%.

Later in 2007, Hanmandlu and Murthy [8] proposed a Fuzzy model based recognition of handwritten Hindi numerals and characters and they obtained 92.67% accuracy for Handwritten Devanagari numerals and 90.65% accuracy for Handwritten Devanagari characters.

In 2008,[9] have used four feature extraction techniques namely, intersection, shadow feature, chain code histogram and straight line fitting features. Shadow features are computed globally for character image while intersection features, chain code histogram features and line fitting features are computed by dividing the character image into different segments. Weighted majority voting technique is used for combining the classification decision obtained from four Multi Layer Perceptron(MLP) based classifier. On experimentation with a dataset of 4900 samples the overall recognition rate observed is 92.80%.

In 2011, [10] proposed an scheme which uses different feature extraction and recognition algorithms. The proposed system assumes no constraints in writing style, size or variations. First the character is preprocessed and features namely: Chain code histogram, four side views, shadow based are extracted and fed to Multilayer Perceptrons as a preliminary recognition step. Finally the results of all MLPs are combined using weighted majority scheme. It is observed that the proposed system achieves 98.16% recognition rates as top 5 results and 89.58% as top 1 results and in [11] proposed a scheme multistage feature extraction and classification scheme The initial stages of feature extraction are based upon the structural features and the classification of the characters is done according to the structural parameters into 24 classes. The final stage of feature extraction employs wavelet transform. Single level wavelet decomposition is used to generate the approximation coefficients which are used as features. The average recognition rate is found to be

92.14% and 94.22% respectively for training and testing samples with wavelet approximation features and 94.68% and 93.23% respectively for training and testing samples with modified wavelet features.

By curvelet transform based approach, [12] the resultant large dimensional feature space is handled by careful application of Principal Component Analysis (PCA). The Support Vector Machine (SVM) and k-NN classifiers were used with one-against-rest class model. Results of Curvelet feature extractor and classifiers have shown that Curvelet with k-NN gave overall better results than the SVM classifier and shown highest results (93.21%) accuracy on a Devanagari handwritten words set.

In [13], presents the application of weighted majority voting technique for combination of classification decision obtained from three Multi Layer Perceptron (MLP) based classifiers for Recognition of Handwritten Devnagari characters using three different feature sets. The features used are intersection, shadow feature and chain code histogram features. Shadow features are computed globally for character image while intersection features and chain code histogram features are computed by dividing the character image into different segments. On experimentation with a dataset of 4900 samples the overall recognition rate observed is 92.16% as we considered top five choices results.

In [14],method using neural network is presented in this paper. Diagonal based feature extraction is used for extracting features of the handwritten Devanagari script. After that these feature of each character image is converted into chromosome bit string of length 378. It is attempted to use the power of genetic algorithm to recognize the character. In first step, preprocessing on the character image, then image suitable for feature extraction as here is used. Diagonal based feature extraction method to extract 54 features to each character. In the next step character.

III. Conclusion

Offline handwritten Hindi character recognition is a difficult problem, not only because of the great amount of variations in human handwriting, but also, because of the overlapped and joined characters. Recognition approaches heavily depend on the nature of the data to be recognized. Since handwritten Hindi characters could be of various shapes and size, the recognition process needs to be much efficient and accurate to recognize the characters written by different users.There are few reasons that create problem in Hindi handwritten character recognition. Some characters are similar in shape (for example भ and म). The character can be written at different location on paper or in window Characters can be written in different fonts. The whole idea behind this research to develop a system that can provide the maximum accuracy and highest recognition rate for character recognition.

REFERENCES

[1] Shruti Agrawal, Dr.Naveen Hemarajani, " Offline Handwritten character recognition with Devanagari Script ",2013.

[2]Ved Prakash Agnihotri ,"Offline Handwritten Devanagari Script Recognition ",2012.

[3] Janhavi Jayant Patil and Dinkar L. Bhombe,"Soft Computiing Approach For Devangari Character Recognition",2013.

[4] Dr. Deepa Gupta, Leema Madhu Nair Improving OCR by Effective Pre-aprocessing And Segmentation For Devangiri Script: A Quantifide Study.

[5] S Ramachandrula, S Jain and H Ravishankar," Offline Handwritten Word Recognition in Hindi", 2012.

[6] Ratnashil N Khobragade, Dr. Nitin A. "A Survey on Recognition of Devnagari Script",2013.

[7] Sandhya Arora, Latesh Malik , Debotosh Bhattacharjee,Mita Nasipuri,"A Novel Approach for Hamdwritten Devangari Charavter Recognition".

[8] M. Hanmandlu, O.V. Ramana Murthy, Vamsi Krishna Madasu, "Fuzzy Model based recognition of Handwritten Hindi characters", IEEE Computer society, Digital Image Computing Techniques and Applications, 2007.

[9] S. Arora, D. Bhattacharjee, M. Nasipuri, D.K. Basu and M.Kundu, "Combining Multiple Feature Extraction Techniques for Handwritten Devnagari Character Recognition",2008

[10] S. Arora, D. Bhattacharjee, M. Nasipuri, D.K. Basu and M.Kundu"Complementary Features Combined in a MLP-based System to Recognize Handwritten Devnagari Character", 2011.

[11] S Shelke and S Apte,"A Multistage Handwritten Marathi Compound Character Recognition Scheme using Neural Networks and Wavelet Features", 2011.

[12] Brij mohan Singh ,Ankush Mittal, M.A. Ansari ,Debashis Ghosh , " Handwritten Devanagari Word Recognition: A Curvelet Transform Based Approach"

[13] Pratibha A Desai, Sumangala N Bhavikattiand RajaShekhar Patil," Neural Networks Based Offline Handwritten Character Recognition System,2013

[14] Dr. Deepa Gupta, Leema Madhu Nair Improving OCR by Effective Pre-aprocessing And Segmentation For Devangiri Script: A Quantifide Study.

[15] Ganesh S. Sable and Sheetal Arun Nirve ,"Optimization of Optical Character Recognition for Printed Devangiri Script using ANFIS Techniques",2013

AUTHORS BIOGRAPHY

Mr. KAPIL BAMNE obtained his B.E degree in ECE from Shri Govindram seksaria institute of technology & science ,indore ,India. He is PG research scholar in Communication Engg from the Government Ujjain engineering college , Ujjain, India. His areas of interests include Microcontrollers, embedded systems, image recognition. Email address: kapil.bamne2187@gmail.com