Handwritten Devnagri Character Recognition

Prof. Shraddha V. Shelke
Assistant Professor
Department of Electronics & Telecommunication Engineering
K. K. Wagh Institute of engineering Education and Research Nashik, Maharashtra

Dr. Prof. Dinesh M. Chandwadkar
Head of Department
Department of Electronics & Telecommunication Engineering
K. K. Wagh Institute of engineering Education and Research Nashik, Maharashtra

Abstract

Recognition of handwritten characters has been a popular research area for many years. Devnagari script is a major script of India and is widely used for various languages. In this paper we propose a system to recognize devnagri handwritten characters. Total 60 devnagri characters (50 letters and 10 digits) are taken in to consideration. 60 samples of each character i.e. total 3600 samples are used for features extraction. Classification is done by four different classifiers which are Multilayer perceptron, K-Nearest Neighbour, Naive Bayes classifier and Classification tree. Performance of different classifiers is compared.98.9 % accuracy is obtained by Multilayer perceptron.

Keywords- Devnagri Characters, Multilayer perceptron, K-Nearest Neighbour, Naive Bayes

I. INTRODUCTION

Recognition of handwritten characters has been a popular research area for many years because of its various applications. Devnagri is the most ancient and perfect among the great languages of the world. Its storehouse of knowledge is an unsurpassed and the most invaluable treasure of the world. Lots of ancient literature is written in Devnagri language. Due to increase in machine dependent applications like translation, language interpretation, sorting mail, reading checks etc. The OCR systems have gained importance in recent years.

Devnagari Character Recognition system is an active yet challenging area of research [1]. With the increasing demand of computers in offices and homes, automatic processing of handwritten paper documents is gaining importance. Devnagari script is used for writing many official languages in India, e.g. Hindi, Marathi, Sindhi, Nepali, Sanskrit and Konkani, also Hindi is the national language of India. Hindi is the third most popular language in world [2].

The attempts were made as early 1977 in a research report on handwritten Devnagari characters [3] with a limited success. Devnagari script is written by joining the characters, even merging characters to have compound characters and also putting "matras" in various forms, this makes the interpretation or recognition extremely difficult. Thus there was very slow progress in the automatic recognition of devnagari script although there are many script and languages in India but not much research has been done for the recognition of handwritten Indian characters. System based on Support Vector Machines (SVM) and Modified Quadratic Discriminant Function (MQDF) for the recognition of off-line handwritten Devnagari characters is suggested by Umapada Pal, Sukalpa Chanda[4].Many techniques have been proposed in the literature for recognizing unconstrained handwritten Devnagari character recognition. The techniques includes: Chain code used by Sharma, N. and U. Pal, structural code used by Arora S., D. Bhatcharjee, gradient and Eigen deformation used by Mane, V. and L. Ragha.

II. STAGES IN CHARACTER RECOGNITION

The functional blocks of Handwritten Character recognition System includes Data Collection, Registration, Preprocessing, Feature Extraction, and Classification as shown in figure 1

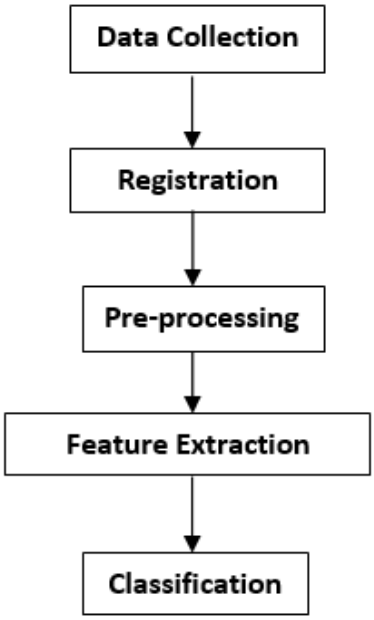

Fig. 1: Stages in character Recognition system

A. Data Collection

The first stage in any pattern recognition system is data collection. Before a pattern is made up of a set of measurement, these measurements need to be performed using some technical equipment and converted to numerical form. In the case of character recognition, such equipment includes video camera, scanners or digital board.

B. Registration

If captured images are in RGB scale. These images have to be converted into greyscale format before further processing. Using appropriate greyscale thresholding; binary images are to be created. So that it will be easy to convert gray images in to binary without loss of information.

C. Preprocessing

Real world input data always, contains some amount of noise. One of the primary reasons preprocessing is to reduce noise and inconsistent data. Noisy data can obscure the underlying signal cause confusion, especially if the key input variable is noisy. Pre-processing can often reduce noise and enhance the signal.

D. Feature Extraction

In feature extraction stage each character is represented as a feature vector, which becomes its identity. The major goal of feature extraction is to extract a set of features, which maximizes the recognition rate with the least amount of elements. Due to the nature of handwriting with its high degree of variability and imprecision obtaining these features, is a difficult task. Feature extraction methods are based on 3 types of features:

1) Statistical
2) Structural
3) Global transformations and moments

Representation of a character image by statistical distribution of points takes care of style variations to some extent the major statistical features used for character representation are:

1) Zoning
2) Projections and profiles
3) Crossings and distances

The character image is divided into NxM zones. From each zone features are extracted to form the feature vector. The goal of zoning is to obtain the local characteristics instead of global characteristics

E. Classification

Features obtained after feature extraction are classified by various classifiers like k-Nearest Neighbour (k-NN) , Bayes Classifier, Neural Networks (NN), Hidden Markov Models (HMM), Support Vector Machines (SVM), etc. There is no such thing as the “best classifier”. The use of classifier depends on many factors, such as available training set, number of free parameters.

III. IMPLEMENTATION

As mentioned in section II data collection can be done by video video camera, scanners or digital board, we used directly available Devnagri characters dataset. Dataset is obtained by collecting handwritten characters from different people from different age groups. Figure 2 shows sample image of character u.

MATLAB code will convert input image of characters in to binary 20X28, 10X14 and 5X7matrix as shown in figure.2. By reshaping matrix elements feature vector is obtained for each character. Number of elements in each feature vector is 560 for 20X28matrix, 140 for 10X14 matrix and 35 for 5X7 matrix. 60 devnagri characters (50 letters and 10 digits) are taken in to consideration.60samples of each character obtained from different users are used for features extraction. Total 60*60=3600samples are used for training purpose.

Fig. 2: Sample image of character u

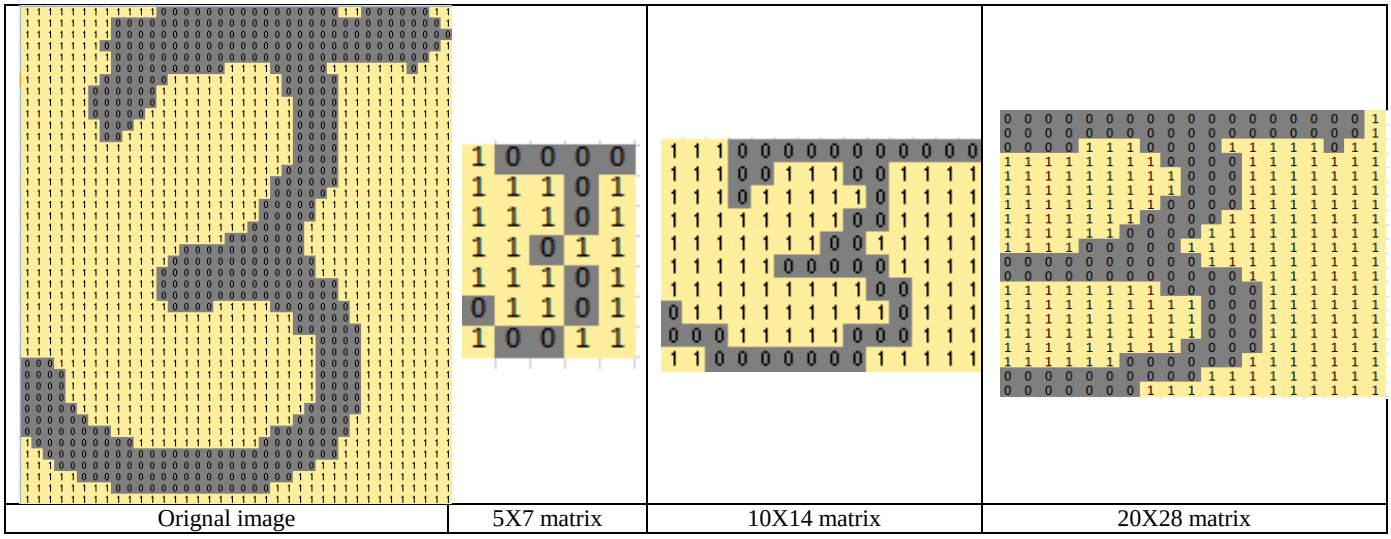

Fig. 3: Binary input image of character u in 5X7, 10X14 and 20X28 matrix.

WEKA 3.6 Machine learning software is used for classification purpose. Advantage of using WEKA is that we can select different classifiers and observe their performance. For WEKA data can be imported from a file in various formats: ARFF, CSV, and C4.5, binary. MATLAB code will directly generate input file in .csv file or .arff file

Following Classifiers are used to classify features

1) Multilayer perceptron.
2) Decision Tree (J48)
3) K-nearest neighbours' classifier. (K=1)
4) Naive Bayes

Figure 4 shows Multilayer perceptron model generated for 5X7 matrix in WEKA

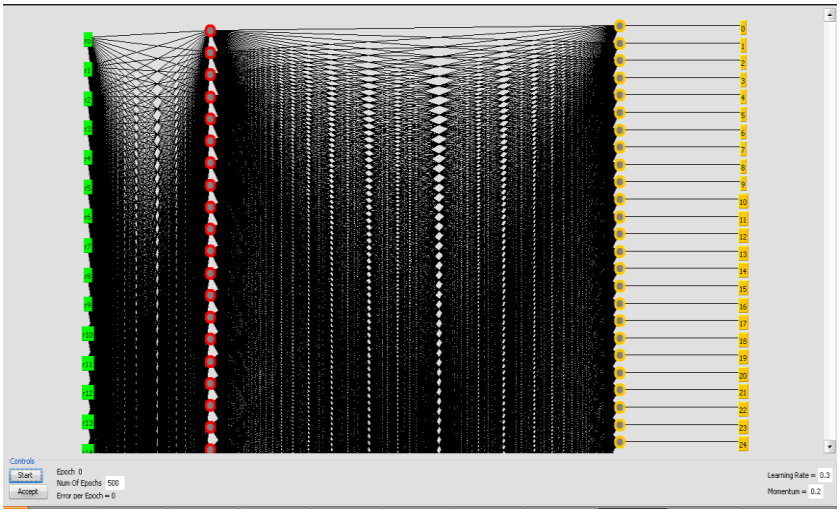

Fig. 4: Multilayer perceptron model

IV. RESULT

Number of correctly classified characters when dataset is used as a training set is given in table 1 and when supplied test set are shown in table 2.
TRAIN SET: 60 samples of each character (60*60 = 3600 instances)
TEST SET: 10 samples of each character (60*10 = 600instances)

Table 1: Classification results of correctly classified characters when dataset is used as training set

Classifier	*Multilayer perceptron*	*K-Nearest Neighbour*	*Naive Bayes*	*Trees J-48*
5X7	*37.5*	*51.5*	*23.7*	*33*
10X14	*85.9*	*98.9*	*57.8*	*72*
20X28	*98.9*	*100*	*70.5*	*79.91*

Table 2: Classification results of correctly classified characters when supplied test set.

Classifier	*Multilayer perceptron*	*K-Nearest Neighbour*	*Naive Bayes*	*Trees J-48*
5X7	*33.66*	*50*	*21*	*27.16*
10X14	*85*	*99.16*	*54.3*	*70*
20X28	*98*	*100*	*63.8*	*80*

V. CONCLUSION

From results we can observe that number of correctly classified characters is more if matrix dimension is increased from 5X7 to 20X28. This is because more the number of elements in feature vectors, accuracy is more. 100% accuracy is obtained by using K-Nearest Neighbour classifier. Classification trees gives accuracy of 80 % and Naive Bayes classifier gives accuracy of 63.8%.Accuracy obtained by Multilayer perceptron is 98%. Time required for training increases from 5 X7 to 20 X28 as number of features increases as it has to process large amount of data. It is also observed that among above classifiers K-Nearest classifiers takes 0.02sec to build model , Naïve Bayes classifier takes 1.16 sec and classification tree take 12.41 sec to build model for 20X28 matrix.

REFERENCES

[1] Jayadevan, R., Satish R. Kolhe, Pradeep M. Patil and Umapada Pal, 2011. "Offline Recognition of Devanagari Script: A Survey", IEEE Transactions on Systems, Man and Cybernetics, Part C, 41(6): 782-796
[2] Pal, U. and B.B. Chaudhuri, 2004. “Indian script character recognition: A survey,” Pattern Recognit., 37: 1887-1899
[3] I.K. Sethi and B. Chatterjee, “Machine Recognition of constrained Hand printed Devnagari”, Pattern Recognition, Vol. 9, pp. 69-75, 1977
[4] Umapada Pal, Sukalpa Chanda,Tetsushi Wakabayashi, Fumitaka Kimura, ”Accuracy Improvement of Devnagari Character Recognition Combining SVM and MQDF”
[5] VelappaGanapathy, and Kok Leong Liew,” Handwritten Character Recognition Using Multiscale Neural Network Training Technique”, World Academy of Science, Engineering and Technology 39 2008
[6] R. Bouckaert, Frank, M.Hall, R.kirkby, P. Reutemann,A. Seewald ,D. Scuse ,”WEKA Manual for Version 3-6-2”, Ch no 4.3 P.no.41,January 11, 2010