Classification of Hindi numeral using Fuzzy Zoning and SVM

Pratibha Singh
Dept. Of Electronics and Instrumentation
Institute of Engineering and Technology,
DAVV Indore (M.P.) India
prat_ibh_a@yahoo.com

Ajay Verma
Dept. Of Electronics and Instrumentation
Institute of Engineering and Technology,
DAVV Indore (M.P.) India
ajayrt@rediffmail.com

Narendra S. Chaudhari
Dept. Of Computer Sc. Engineering
Indian Institute of Technology,
Indore (M.P.) India
nsc183@gmail.com

Abstract— Zoning is widely used for obtaining topological features of handwritten characters. In this paper we have modified the zoning criteria using fuzzy boundary of individual zones keeping the point in mind that writing style of different persons may shift the structural pattern of the character from one zone to the adjacent zone. The zone-wise normalized distance of individual black pixel along with the angle is used as feature. The SVM with radial Basis function as kernel is used for classification.

Keywords— Multi layer Perceptron, support vector machines, Radial Basis Function, Fuzzy logic, Neural network

I. INTRODUCTION

The need to recognize the handwritten text is challenging problem not only from the perspective of behavioural biometrics but also in the context of pattern recognition. Although many pieces of work had been done for recognition of roman script but only few attempts have been tried for Indian scripts. Hindi is the most popular language of India which is written and encoded using Devnagri script.

There are two fundamental approaches to character recognition: Feature based classification and template matching. Template matching based approach is sensitive to size and style variation therefore we go for the first one i.e. feature based classification.

A key reason for the absence of sustained research efforts in Devnagri OCR appears to be the paucity of data resources. Ground truthed data for words and characters, on-line dictionaries, corpora of text documents and reliable, standardized statistical analyses and evaluation tools are currently lacking. So, the creation of such data resources will undoubtedly provide a much needed fillip to researchers working on Devnagri OCR.

A. Introduction to Devnagri script

Devnagri is most popular script of India. Hindi, Sanskrit, Marathi and Nepali is written in Devnagri script. Hindi is most popular language of India and the third most popular language of the world.

The basic characters of modern Devnagri script 14 vowels and 37 consonants. The writing style is from left to right. The concept of lowercase and uppercase is absent. Vowel following consonant takes modified shape known as modified character. The consonant following consonant forms compound character. We focused our research on the Hindi

isolated numeral, whose dataset is obtained from CVPR unit, ISI Kolkata.

II. RELATED LITERATURE

Many pieces of work have been done towards the recognition of Indian printed characters at present OCR systems are commercially available for some of the printed Indian scripts [3]. Although several pieces of research work exist on Indian printed characters but only a few attempts have been made towards the recognition of Indian off-line handwritten characters [2]. Among offline Indian scripts maximum research has been done for Bangla for numeral, characters and unconstrained word recognition [5, 6]. First research report on Devnagri script was published in 1977[4] but not much work is done after that.

Some research reported recently on handwritten Devnagri characters. Hamandlu and Murthy [7] used fuzzy model based recognition for handwritten Hindi numerals, where they used Normalized distance as feature for individual boxes and obtained 92.67% accuracy. Pal et al [6] proposed lexicon driven approach for handwritten unconstrained Bangla text. They used water reservoir principle for obtaining primary primitives from Bangla text which are then merged to form optimum primitives using dynamic programming approach. Directional chain codes are calculated for blocks in bounding boxes which are applied to MQDF based classifier for obtaining total likelihood of the character. Ramteke et al [8] proposed handwritten isolated Marathi numeral recognition scheme based on invariant moments. They used Gaussian distribution function for classification and obtained 87 % accuracy. Bajaj et al [9] employed set of different features namely density, moment and descriptive component feature for classification of Devnagri numerals. They used multiclassifier connectionist architecture for increasing reliability and obtained 89.6% accuracy. Kumar and Singh [10] proposed Zernika moment feature based approach for Devnagri character recognition and artificial neural Network based classifier. Sethi and Chatterjee [4] proposed a decision tree based approach for constrained hand printed Devnagri characters using primitive features.

Bhattacharaya et al [11] proposed a multilayer perceptron (MLP) neural network based classification approach for Devnagri numerals and obtained 91.28 % results. They used multiresolution features based on wavelet transforms. Sharma
et al [12] proposed quadratic classifier based approach for Devnagri character recognition and obtained 98.86 % accuracy for numerals and 80.36 % for characters. Rajashekararadhya et al [13] proposed an efficient zone based feature extraction method for numeral recognition of four popular south Indian scripts. They used Zone centroid and image centroid based distance metric for feature extraction. The above method can be extended to larger data set.

Deshpande et al [15] proposed a regular expression based approach to Devnagri character recognition for increasing efficiency. In this approach the characters are first converted into sequences and then geometrical properties of characters are converted into regular expression which is used for matching purposes. Now the regular expression is tested with minimum edit distance. Bishnu et al [16] proposed Recursive contour following algorithm for character segmentation in which he has divided the word into three zones M, B, E and to find character extent he applied recursive contour following between B & E regions. But this method fails when the character touches the zone of contour following.

Graves et al [17] proposed a novel approach for unconstrained handwriting. They used a novel type of recurrent neural network; specifically designed for sequence labelling tasks where the data is hard to segment contains long-range bidirectional interdependencies. They achieved word recognition accuracies of 79.7 percent on online data 74.1 percent on offline data. Pal et al proposed a [19] segmentation scheme for unconstrained Bangla script in which he divided the document into vertical strips. The horizontal histogram of the strips are obtained, minima of histogram forms the basis for text line separation and vertical projection profile for word segmentation. Character segmentation is done using water reservoir principle. Majumdar et al [20] proposed MLP based classification scheme for the handwritten as well as printed Bangla numeric. In their scheme they used structural features like the position of holes, intersection points, terminal points, curves, average Intensity of individual zone and obtained 97 % accuracy.

Pal et al [23] proposed a scheme for the separating machine printed & handwritten text as their OCR systems will be different. He used longest run length of the characters as first level feature & Head line deletion forming second basis for separating machine printed text from handwritten text. Sural et al [24] set of fuzzy features on Hough transform of character pattern pixel from which additional fuzzy sets are synthesized using t- norm. MLP’s are trained with these sets to recognize the characters. In [25] Siddiki et al used chain code histogram, and obtained first & second order differential chain code for writer identification. Pal et al [27] proposed a scheme for multi oriented text extraction from printed artistic documents. In the proposed scheme by them they used water reservoir principle. Chaudhari et al [29] presented a pioneering effort for the development of handwritten numeral database of Indian scripts. In [30] Bhattacharyya et al proposed a hybrid system for Bangla handwritten numeral recognition based on partially labeled two-layer SOM and MLP classifiers. Hybrid system achieves 96.7% correct recognition rate.

III. PROPOSED METHODOLOGY

Selection of appropriate features is very important for good performance of recognizer. Because we know that human reasoning is some what fuzzy in nature which enables us to recognize even degraded patterns. In this paper we propose a zoning based feature extractor whose boundary is not sharply defined but it is fuzzy.

A. Pre-processing

A set of pre-processing steps were applied to images in order to remove noise and simplify the procedure of feature extraction.

Binarization: The greyscale images stored in tif format is binarized using histogram based global thresholding algorithm. A 2 by 2 median filtering is applied on binary image to remove isolated pixels.

Normalization: The bounding box of each binary image is obtained. Then the image is corrected for slant after which it is resized to standard size of 72x54.

Thining: The skeleton of the resized image is obtained using morphological skeletonization. Then the image is subjected to

Fig. 1: Flow Chart of Proposed Methodology

Fig.2: Raw Images of the Hindi Digits
Proc. of the International Conference on Advanced Computing and Communication Technologies (ACCT 2011)

spur removal algorithm. Now the image is ready for the next step i.e. features extraction.

(a) Original image    (b) Binary noisy image

(c) Image after filtering   (d) image of thinned image

Fig. 3 Output of Different Stages of Pre-processing

IV. FEATURE EXTRACTION

The entire bounding box of the image is divided into 3x3 overlapping zones. The boundary of each individual zone is not sharply defined and the pixels falling in the zones are assigned the membership values such that those who belong to zone i (where i vary from 1 to 9) is assigned membership value 1. And the pixels falling in overlapping area of zone are assigned values 0.75, 0.5 and 0.25 respectively according to the belongingness.

After defining the membership value, the feature vector made up of the normalized distance and the normalized angle of each black pixel is calculated which is then multiplied by corresponding membership value to obtain eighteen element vector of distance, angle of nine zones.

TABLE I
MEMBERSHIP VALUES IN ZONES DEFINED IN FIGURE 1.

<table>
<thead>
<tr>
<th>Sub-zone</th>
<th>Membership Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zone-1</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>0.75</td>
</tr>
<tr>
<td>B</td>
<td>0.5</td>
</tr>
<tr>
<td>C</td>
<td>0.25</td>
</tr>
<tr>
<td>Zone-2</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>0.25</td>
</tr>
<tr>
<td>B</td>
<td>0.5</td>
</tr>
<tr>
<td>C</td>
<td>0.75</td>
</tr>
</tbody>
</table>

Note that the pixels of the sub-zone 1 lying in region A, B, C are assigned membership values 0.75, 0.5, 0.25 respectively. While the same regions defined for sub-zone 2 have the values 0.25, 0.5, 0.75 respectively.

V. CLASSIFICATION

Classification is performed using SVM. SVMs use an implicit mapping $\Phi$ of the input data into a high-dimensional feature space defined by a kernel function, i.e., a function returning the inner product ($\Phi(x), \Phi(x')$) between the images of two data points $x$, $x'$ in the feature space. The learning then takes place in the feature space. Here we have used RBF kernel. This kernel nonlinearly maps samples into a higher dimensional space so it, unlike the linear kernel, can handle the case when the relation between class labels and attributes is nonlinear.

SVM is a binary classifier. For our multiclass problem ‘one against all’ strategy is used where we have trained 10 binary SVM classifiers. The classifiers are then combined by comparing their decision values on a test data instance and labeling it according to the classifier with the highest decision value.

VI. EXPERIMENTAL RESULTS

The data samples of handwritten Hindi numerals were obtained from CVPR unit, ISI Kolkata. The dataset consists of 200 grayscale image of each class. After preprocessing angle and normalized distance feature is calculated. The feature vector size is eighteen which is made up of normalized angle and normalized distance of each of the nine zones. The 5-fold cross validation is used to generate the test set. The Performance of the recognizer is measured in terms of correct rate and error rate which are defined as:

Correct rate = \[ \frac{\text{Correctly Classified Samples}}{\text{classified samples}} \]

Error Rate = \[ \frac{\text{Incorrectly Classified Samples}}{\text{Classified Samples}} \]

TABLE II
PERFORMANCE OF CLASSIFIER.

<table>
<thead>
<tr>
<th>Hindi Numeral</th>
<th>Correct rate</th>
<th>Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zero</td>
<td>0.9925</td>
<td>0.0075</td>
</tr>
<tr>
<td>One</td>
<td>0.9450</td>
<td>0.0550</td>
</tr>
<tr>
<td>Two</td>
<td>0.9075</td>
<td>0.0925</td>
</tr>
<tr>
<td>Three</td>
<td>0.9275</td>
<td>0.0725</td>
</tr>
<tr>
<td>Four</td>
<td>0.9150</td>
<td>0.0850</td>
</tr>
<tr>
<td>Five</td>
<td>0.9175</td>
<td>0.0825</td>
</tr>
<tr>
<td>Six</td>
<td>0.9350</td>
<td>0.0650</td>
</tr>
<tr>
<td>Seven</td>
<td>0.9500</td>
<td>0.0500</td>
</tr>
<tr>
<td>Eight</td>
<td>0.9075</td>
<td>0.0925</td>
</tr>
<tr>
<td>Nine</td>
<td>0.9275</td>
<td>0.0725</td>
</tr>
</tbody>
</table>

VI. CONCLUSIONS

In this paper a new method based on zoning with fuzzy boundary was presented. The proposed method is
easy to implement as well as gives satisfactory performance for limited training sample. The main advantage of the proposed scheme is the length of feature vector. The less size of feature vector reduces the computation time as well as giving satisfactory performance in recognizing Hindi digits. So, it may be used practically because practically a large number of database gathering is not feasible.

ACKNOWLEDGMENT

We are very thankful to the Director, IET, DAVV, Indore for his constant support and motivation.

REFERENCES