Zone Based Feature Extraction Techniques For Bangla Numerals Recognition

Gita Sinha  
Research scholar  
AISECT university Bhopal  
gitawit321@gmail.com

Biresh Kumar  
Research scholar  
NIT Patna  
bireshmtech@gmail.com

Abstract
This paper proposed a methodology for Handwritten Bangla numerals Recognition using zone based Feature Extraction Techniques. Every numeral image is pre-processed, segmented and feature are Extracted from each zone. In this paper we present, three zone based feature extraction techniques which is namely: image Centroid zone(ICZ), zone centroid zone (ZCZ) and hybrid feature extraction techniques which is combination of (ICZ+ZCZ).Recognition of the numerals is completed through combining SVM classifier respectively corresponding to the three types of features. The recognition rate of this method is 97.21% on handwritten bangle numeral database.

Key term-
Image centroid zone(ICZ), Zone centroid zone (ZCZ) and support vector machine (SVM).

I. Introduction
Handwritten Numeral Recognition (HNR) is the process of classifying written Numeral into appropriate classes based on the features extracted from each Numeral. Handwritten character recognition can be performed either online or offline. A system has been developed for offline HNR of Bangla writing systems using Support Vector Machines(SVMs). The character set of Indian languages is large and consists of more complex characters when compared to the Latin script. A handwriting or speech based interface to the computer has become imperative for Indian languages because of complex keyboard mapping procedures for Indian language characters. As many Indian languages have a similar character set, developing a recognition system for one Indian language serves as a framework for others as well.

Most character recognition applications target machine printed and handwritten characters on paper documents. Recently, the recognition of text in videos, web documents, and natural scenes has become an urgent demand because this task is difficult to realize [1]. OCR consists with several steps namely pre-processing, segmentation, feature extraction and classification which describe latter on this paper. OCR work on printed Bangla Character started in early 1990s [2] and no significant work is reported till mid-1990s. Recently, several pieces of work on Bangla have been published [3]. Among the earlier pieces of work, some of the efforts on Bangla character recognition are due to Ray and Chatterjee[4]. The first complete system capable of doing OCR from printed Bangla documents is due Chaudhari and Pal [7]. In this system pre-processing involves skew correction, followed by noise removal, and preliminary segmentation of the input image into lines, zones and characters.

In 2008 Mohammed Moshiulet al. [5] They proposed a methodology for recognizing Bangla handwritten numerals which are based on fuzzy logic in this paper, they used unique fuzzy rule base for each numeral and tested the engine for Bangla numerals considering various writing style and got more than 80% recognition accuracy.

In 2011 Poulami Das [6] have proposed a novelscheme for recognition of offline basic characters of Bangla using multiple classifiers. they have adopted hybrid approach because it is nearly impossible to find a set of stroke features which are sufficient to classify the characters. A prototype of the system is tested with a data set containing 4423 characters of different font and size. On average, the recognition accuracies for Binary tree based classifier and Multilayer perceptron [with backpropagation for learning] (MLP) are 90% above approximately.

In 2005, K. Roy et al. [8] They have used directional features extracted from the contour of each numeral to extract these features, the minimum bounding box containing the image of each numeral is first segmented into few blocks and then the direction code histogram is computed with each of these blocks. Peak values of each such histogram are considered as the feature values of the corresponding blocks. Considering all the blocks a 100 element feature set is formed for representation of each image pattern and a database of 12000 numerals are used for the same, they achieve 96.93% of accuracy. They have used database consists of 600 samples of online handwritten Bangla numerals. In each class there are 600 samples. They achieve 96.59% and 95.86% of accuracy.

In 2009, Saleh Ahmed [9] this paper presents Bangla numeral Character Recognition System using supervised locally linear embedding algorithm and Support Vector Machine (SVM). The locally linear embedding (LLE) algorithm is an unsupervised
A technique proposed for nonlinear dimensionality reduction.

II. Steps in OCR

Input images

- Digitisation
- Pre-Processing
- Segmentation
- Feature Extraction
- Classification

Fig-1 stages in OCR

Pre-processing

This is the step taken before the major image processing task. The problem here is to perform some basic tasks in order to render the resulting image more suitable for the job to follow. Image pre-processing can significantly increase the reliability of an optical inspection. It may involve enhancing the contrast, removing noise, Normalization, Binarisation, Filtration and Smoothing.

- Enhancing Contrast - Image contrast enhancement based on a generalization of histogram equalization (HE). It is a useful technique for improving image contrast. [10]

- Removing noise - Noise reduction is the process of removing noise from an image. Noise can be random or white noise with no coherence, or coherent noise introduced by the device's mechanism or processing algorithms.

- Normalization - The character image is normalized into a window of size 100x100. After normalization, we produce bitmap image of normalized image. Now, the bitmap image is transformed into a contour image.

- Binarisation - Image Binarization refers to the conversion of a gray-scale image into a binary image. Each pixel in an image is converted into one bit and the value ‘1’ or ‘0’ assigned depending upon the value of the pixel and threshold value. Binarization process is categorized into two groups global and local.

- Filtration - Pre-processing methods use a small neighbourhood of a pixel in an input image to get a new brightness value in the output image. Such pre-processing operations are also called filtration.

- Smoothing - Smoothing suppresses noise or other small fluctuations in the image. Calculation of the new value is based on averaging of brightness values in some neighborhood O.

Segmentation

Image segmentation is the process of assigning a label to every pixel in an image such that pixels with the same label share certain visual characteristics. The goal of segmentation is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyze. Image segmentation is typically used to locate objects and boundaries in images. We have used three types of Segmentation process line segmentation, word segmentation and character segmentation.

Feature Extraction

During or after the segmentation procedure the feature set, which is used in the training and recognition stages. Feature sets play one of the most important roles in a recognition system. A good feature set should represent characteristic of a class that helps distinguish it from other classes. These are the some feature Extraction techniques Global Transformation and Series Expansion Features, Statistical Features, Geometrical and Topological Features.

Classification

Classification aims at creating a map from the input data to a corresponding known output in the training phase. The constructed map, called classifier, is then used to predict new input instances. Many classification techniques have been developed such as linear discriminate functions, K-nearest neighbours, Bayesian decision theory, neural networks, committee machine, Support Vector Machine (SVM) and so on. However, all those techniques train and classify data considering only the physical features (e.g., distance or similarity).
Post-processing
In post-processing we want to replace the input sequence of characters with another sequence of characters that is graphically similar and form the likeable sentence of the given language. We also reduce the number of errors in text. Semantic relations between words can be used to aid selection from alternative candidate words output from an Optical Character Recognition (OCR) system in order to improve the overall recognition rate. One method of automatically identifying the semantic relations between words is by using an existing knowledge source [11].

IV. Dataset collection
Bangla dataset consists of 12000 samples of handwritten Bangla numerals. Each class consists of 1200 samples of each numeral. Maximum data has been obtained from [12] and remaining has been contributed by informal communication with people belonging to Bengal. Dataset consists of 10 different folders containing a collection of images. These numerals are shown in table 1 of this chapter and Arabic numerals are figured out in chapter 1.

<table>
<thead>
<tr>
<th>Table 1 Handwritten Bangla Numeral Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
</tr>
<tr>
<td>------------------------------</td>
</tr>
<tr>
<td>02</td>
</tr>
<tr>
<td>02</td>
</tr>
<tr>
<td>02</td>
</tr>
<tr>
<td>02</td>
</tr>
<tr>
<td>02</td>
</tr>
<tr>
<td>02</td>
</tr>
</tbody>
</table>

Feature Extraction
Following listed features have been used for current experiment. Two types of features namely image centroid zone and zone centroid zone. 200 feature vectors have been formed using combinations of both basic features. These methods provide the ease of implementation and good quality recognition. In the next section, these algorithms have been defined. The following paragraph explains the details about feature extraction method.

Image Centroid Zone
The centroid of image (numeral/character) has been computed. The given image has been further divided into 100×100 equal zones where size of each zone is (10×10). Then, the average distance from image centroid to each pixel present in the zones/block has been computed. 100 feature vectors of each image are thus obtained. Zones which are empty are assumed to be zero. This procedure is repeated for all zones present in image (numeral/character).

Figure 3 shows example of Bangla Numeral image(4) of size 32×32. First, centroid of image is computed. Then, image is divided into 16 equal zones each of size 8×8. Later, average distance from image centroid to each pixel present in the image is computed.

Zone Centroid Zone
In ZCZ, image is divided into 100×100 equal zones and centroid of each zone is calculated. Followed by computation of average distance of zone centroid to each pixel present in zone. Zones which are empty are assumed to be zero. This procedure is repeated for all pixels present in each zone. Efficient zone based feature extraction algorithm has been used for handwritten numeral recognition of four popular south Indian scripts as defined in[13] Here, same method has been applied on few north Indian scripts. Algorithm 1 provides Image centroid zone (ICZ) based distance metric feature extraction system, while Algorithm 2 provides Zone Centroid Zone (ZCZ) based Distance metric feature extraction system. Further, Algorithm 3 provides the combination of both (ICZ+ZCZ) feature extraction systems. The following algorithms illustrate the working procedure of feature
extraction methods as depicted in figure 3

Figure 3(ICZ) Image 32×32 and block 8×8 of handwritten Bangla Numerals “Four”

Algorithm 1: Image Centroid Zone (ICZ) feature extraction method.

Input : Pre-processed Image (character/numeral)
Output : Extract the Features for Classification and Recognition.

Method Begins
Step 1: Calculate centroid of input image.
Step 2: Division of input image into 100×100 equal zones.
Step 3: Computation of the distance from the image centroid to each pixel present in the zone.
Step 4: Repeat step 3 for the entire pixel present in the zone/boxes/grid.
Step 5: Average distance computed between these points.
Step 6: Repeat this procedure sequentially for the entire zone present in the image.
Step 7: Obtaining 100 such feature for classification and recognition.

Ends.

Algorithm 2: Zone Centroid and Zone (ZCZ) based feature extraction system.

Method Begins
Step 1: Division of input image into \( n \) equal zones.
Step 2: Compute centroid of each zone.
Step 3: Compute the distance between the zone centroid to each pixel present in the zones.
Step 4: Repeat step 3 for the entire pixel present in the zone/box/grid.
Step 5: Computation of average distance between these points present in image.
Step 6: This procedure are sequentially repeat for the entire zone.
Step 7: Obtaining 100 such features for classification and recognition.

Ends.

Hybrid Algorithm 3: Hybrid feature extraction method is a combination of both of the algorithm (ICZ + ZCZ) defined above. This method provides 200 such features from each of the image.

In figure 5 image size is taken as 50×50. When this image is divided into equal zone of size 10×10 pixels, then total number of zones formed is 25. Similarly image size is taken as 100×100, and image is divided into equal zone of size 10×10 pixels then total number of zones will be 100. Likewise, when image size is 32×32, and block size is 8×8, then 16 zones will be created.
Classification
SVM classifiers have been used in this implementation. SVM is popular & efficient and also produce most efficient results in this implementation.

Support Vector Machines (SVM)
Support Vector Machines (SVM) has been used for the purpose of Classification and Recognition. Support Vector Machines are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships. A Support Vector Machine (SVM) is a concept in statistics and computer science for a set of related supervised learning methods that analyze data and recognize patterns, used for classification and regression analysis. It has capability of learning to achieve good generalization performance, which is objective of any machine, given a finite amount of training data. Striking a balance between goodness of fit obtained on a given training dataset and the ability of machine to achieve error free recognition on all the dataset. The standard SVM takes a set of input data and predicts two possible classes of input. SVM training algorithm builds a model that assigns new examples into one category or the other. SVM utilized in pattern recognition is to construct a hyperplane as the decision plane, which separates the positive and negative patterns with the largest margin. SVM has proved to achieve good generalization performance by the use of concept of basis, without knowledge of the prior data [14].

Cross-Validation
There are two parameters for an RBF kernel: C and γ. It is not known before hand which C and γ are best for a given problem; consequently some kind of model selection (parameter search) must be done. The goal is to identify good (C), so that the classifier can accurately predict unknown data (i.e. testing data). Note that it may not be useful to achieve high training accuracy (i.e. a classifier which accurately predicts training data whose class labels are indeed known). As discussed above, a common strategy is to separate the data set into two parts, of which one is considered unknown. The prediction accuracy obtained from the “unknown” set more precisely reflects the performance on classifying an independent data set. An improved version of this procedure is known as cross-validation. In five-fold cross-validation, First of all, training set is divided into five subsets of equal size. Sequentially one subset is tested using the classifier trained on the remaining 5-1 subsets. Thus, each instance of the whole training set is predicted once so the cross-validation accuracy is the percentage of data which are correctly classified.

Experimental Results
Bangla Numeral Recognition with SVM
Table 1 shows result on different parameter. When the value of parameter increases then accuracy slightly increases. The highest recognition accuracy on feature vectors 2 is 98.45%.

Table 1 Recognition accuracy on different value of C and γ.

<table>
<thead>
<tr>
<th>Sr. No.</th>
<th>Feature vector</th>
<th>γ(gamma)</th>
<th>C</th>
<th>Recognition accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Fv1</td>
<td>0.001</td>
<td>1</td>
<td>96.97%</td>
</tr>
<tr>
<td>2.</td>
<td>Fv2</td>
<td>0.002</td>
<td>2</td>
<td>92.88%</td>
</tr>
<tr>
<td>3.</td>
<td>Fv3</td>
<td>0.01</td>
<td>4</td>
<td>98.39%</td>
</tr>
<tr>
<td>4.</td>
<td>Fv1</td>
<td>0.02</td>
<td>8</td>
<td>97.24%</td>
</tr>
<tr>
<td>5.</td>
<td>Fv2</td>
<td>0.008</td>
<td>16</td>
<td>95.56%</td>
</tr>
<tr>
<td>6.</td>
<td>Fv3</td>
<td>0.001</td>
<td>32</td>
<td>98.45%</td>
</tr>
<tr>
<td>7.</td>
<td>Fv1</td>
<td>0.0016</td>
<td>64</td>
<td>97.22%</td>
</tr>
<tr>
<td>8.</td>
<td>Fv2</td>
<td>0.016</td>
<td>128</td>
<td>97.67%</td>
</tr>
<tr>
<td>9.</td>
<td>Fv3</td>
<td>0.1</td>
<td>256</td>
<td>98.45%</td>
</tr>
<tr>
<td>10.</td>
<td>Fv3</td>
<td>0.1</td>
<td>512</td>
<td>97.22%</td>
</tr>
<tr>
<td>11.</td>
<td>Fv3</td>
<td>0.016</td>
<td>512</td>
<td>98.45%</td>
</tr>
</tbody>
</table>
Another direction we would like to apply these features to achieve better recognition rates for the recognizer. Combined with hybrid feature extraction, and motivate [15].

Figure 6 show Bangla numeral accuracy on fixed value of $\gamma=0.001$ and different value of $C$.

VII. Conclusions
In this paper, a feature extraction method for increasing the rates of recognition of handwritten bangle numerals by combining ICZ + ZCZ. The experiments demonstrated that the combination of (ICZ + ZCZ) with SVMs experts pairs of classes that constitute the greatest improved performance in terms of recognition rate. The results showed a significant improvement from 96.97% to 98.45% in recognition rate for all cases tested on different feature vector.

We have obtained good recognition rate for the numeral using these two(ICZ + ZCZ) feature extraction method, compared to the result obtained using Image Centroid Zone and Zone Centroid Zone. Results are compared to prove the accuracy of the final system along with specific features and its values. There are several factors upon which the accuracy of the system depends such as number of feature vector, values of C and $\gamma$. This approach is more efficient and robust than using a single feature extraction techniques. Our future work aims to improve classifier to achieve still better recognition rate and also to develop new and more accurate algorithms, which provides efficient result, and intend to used in different classifier combined with hybrid feature extraction, and motivate to achieve better recognition rates for the recognizer. Another direction we would like to apply these feature extraction techniques on bangle character recognition [15].

VIII. Acknowledgment
We are very thankful to ISI Kolkata and remaining dataset has been contributed by informal communication with people belonging to Bengal, for his sincere help towards the provision of collected Bangla Numeral dataset for our experiment.

References
1. Minoru Mori, Minako Sawaki, Junji Yamato “Robust Character Recognition Using Adaptive Feature Extraction” 9781-4244,2582-2008 IEEE.
6. Poulami Das “Recognition of Bangla Basic Characters using Multiple Classifiers”, 978-1-4577-1386-2011 IEEE.


15. Washington W. Azevedo and CleberZanchettin “A MLP-SVM Hybrid Model for Cursive Handwriting Recognition” Proceedings of International Joint Conference on Neural Networks, San Jose, California, USA, July 31 – August 5, 2011