ISSN No. 0976-5697

Volume 4, No. 9, July-August 2013

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Support Vector Machines for Odiya Handwritten Numeral Recognition

Rasmi Ranjan Das
Department of Computer Science & Engineering
Gandhi Institute for Technological Advancement,
Badaraghunathpur, Madanpur, Odisha, India - 752054
rasmi82@gmail.com

Swati Sucharita Das
Department of Computer Science & Engineering
Centurion Institute of Technology,
Bhubaneswar, Odisha, India - 752050
das.swati16@gmail.com

Shom Prasad Das
Department of Computer Science & Information Technology Engineering
National Institute of Science and Technology,
Palur Hills, Berhampur, Odisha, India - 761008
shomdas@yahoo.com

Abstract: Handwritten Numeral recognition has its own importance over various fields like automatic pin code reading, bank check processing, form processing, etc. Handwriting numeral recognition is a subset of character recognition. But handwriting recognition has been always a complex task due to variations of handwriting shape, size, and stroke among different writers. This paper focuses on automatic offline recognition of isolated odiya handwritten numerals recognition using two features extraction techniques - Fourier Descriptors and Normalized Chain Code. Once the features are extracted for the numerals, the features are fed into two different machine learning technique i.e., Back Propagation Neural Network (BPN) and Support Vector Machines (SVM) for classification and recognition. A total set of approximately 20000 handwritten numerals data from 200 different classes of people are collected and are considered for the experimentation of the proposed methods. The classification results from the two machine learning techniques are then compared. We have obtained a recognition accuracy of 83.33% & 93.4% with fourier descriptor feature extraction techniques for BPN and SVM respectively. In case of normalized chain code feature extraction technique we obtained 83.63% and 93.57% for BPN and SVM respectively. The experimented result shows SVM outperform the BPN with both the feature extraction techniques for recognizing and classifying odiya numeral.

Keywords: Handwritten Numericals, Feature Extraction, Fourier Descriptors, Normalized Chain Code, Back Propagation Neural Network, Support Vector Machines.

I. INTRODUCTION

Handwriting recognition has always been a challenging task .It is the ability of a computer to receive and interpret intelligible handwritten input from sources such as paper documents, photographs, touch screens and other devices. The recognition system can be either online or off-line. In on-line handwriting recognition, words are generally written on a pressure sensitive surface (digital tablet PCs) from which real time information, such as the order of the stroke made by the writer is obtained and preserved in the other hand off-line handwriting recognition is the process of finding letters and words that are present in the digital image of a handwritten text.

This paper represents methods of recognition of the odiya numerals. Odiya is an Eastern Indo-Aryan language belonging to Indo-Aryan language family. It is mainly spoken in the Indian state of Odisha. The script has 49 characters and 10 different odiya numerals of the decimal number system.

Rest of the paper is organized as follows. Section 2 gives a brief overview of the two machine learning techniques used in our research for recognition and classification. In Section 3 some related works are presented followed by data collection and pre processing techniques in section 4. Section 5 discusses the algorithms used for feature extraction and in Section 6 experimental results are discussed. Section 7 focuses on conclusion of the research.

II. OVERVIEW OF BPN AND SVM FOR CLASSIFICATION

Back propagation neural network (BPN) is typical network consisting of set of sensory units that constitute the input layer, one or more hidden layers of computation nodes and an output node. All the neurons of one layer are fully interconnected with all neurons of its just preceding and just succeeding layers. These neural networks are also referred to as multilayer perceptrons (MLPs). BPN are applied to solve some difficult diverse problems by training them in a supervised manner with a highly popular algorithm known as error back-propagation algorithm. This algorithm is based on error correction learning rule. Basically the error propagation algorithm consists of two passes: a forward pass and a backward pass. In forward pass the input vector is applied at the nodes of the network and effect propagates through the network layer by layer and finally output is produced as the actual response of the system. The synaptic weight during the forward pass process is fixed but during the backward pass the synaptic weights are adjusted in accordance with an error correction rule. The error signal is transmitted backward through the network. Hence the network is called as the back propagation neural network [1].

Support Vector Machines (SVM) has been considered as one of the powerful classifiers for character and numeral recognition. Basically it is a linear machine with the main idea of constructing a hyperplane as the decision surface in a way that the margin of separation between two classes is maximized. The main objective is to do better classification. The SVM is an approximate implementation of the method of structural risk minimization. This induction principle is based on the fact that the error rate of a learning machine on test data is bounded by the sum of the training error rate and a term that depends upon the Vapnik-Chervonenkis(VC) dimension. Accordingly, the support vector machines can provide a generalization performance on pattern classification problems despite the fact that it does not incorporate problem-domain knowledge. A notion that is central to the construction of the support vector learning algorithm is the inner product kernel between a support vector X(i) and the vector X drawn from the input space. The support vectors consist of a small subset of training data extracted by the algorithm [1].

We may use the support vector learning algorithm to implement the following types of learning machines:

polynomial learning machines, radial-basis function networks and two layer perceptrons(i.e; with a single hidden layer)

III. RELATED WORK

In past, various works have been carried out by researchers for offline handwritten numeral recognition of many foreign and other Indian languages. Every work has been based upon different approaches and techniques of recognition. In [2 and 3] the image centroid and zone based feature extraction technique is proposed for the recognition of four popular south Indian scripts. In this approach the centroid of the numeral is computed and image is divided into N equal zones. Then the average distance from the centroid to each pixel in the zone is computed and is repeated for all the zones. Finally N such features are extracted for classification and recognition. In this approach the average recognition rate obtained are observed as 96% for Tamil and Malayalam numerals and about 98% for Telugu numerals.

In [4] study has been done regarding the fourier descriptors and chain code of the image to extract features. Fourier transform is computed at the boundary of the image and features are extracted. In chain code technique freeman chain code are generated from the boundary of the image which are then normalized to get the feature vector. The feature extracted is then used for the classification and recognition through the SVM classifier for the recognition of Marathi handwritten numerals. This paper achieved an overall recognition rate of 98%.

In [5] , the paper dealt with recognizing the handwritten numerical strings which are more complex than recognizing isolated digits since it deals with problems such as segmentation, overlapping etc. This paper uses a segmentation based recognition system using heuristic over-segmentation. The paper is done in two folds firstly by SVM and then by multi layer perceptron neural networks. The results achieved were about 99% for both the fold.

In [6], it dealt with classification through SVM for recognizing the handwritten numerals for Devanagari scripts. Moment invariant and affine moment invariant techniques are used as feature extractor. Based on normalized central moments, asset of seven moment invariants is derived and the resultant was thinned and seven moments were extracted. Thus 14 features are extracted again by affine invariant moment method by means of theory of algebraic invariants. Thus 18 features are collected to be used in SVM classifier for recognition and classification and an overall recognition rate of 99.48% was obtained.

In [7], to care of huge variability and distortion of patterns off-line unconstrained Malayalam handwritten numeral is proposed. Topological and structural features are used for recognition. It's mainly done to illustrate the water-reservoir concept. Close loop features are the main topological features used here and in structural feature morphological pattern is considered. An overall recognition rate of 96% was obtained in this method.

In [8], it dealt about the Malaysian car plate recognition with the chain code techniques to perform recognition of different types of fonts used in Malaysia. The characters are recognized by using the list of chain codes derived for each character from the previous phase. Here the total of each code is used as a guide to recognize the characters by matching the chain codes extracted from the previous phase. It has obtained an average recognition accuracy of 76.19%.

In this paper we have implemented fourier descriptor and normalized chain code techniques for feature extraction from the Odiya numerals. We have selected Odiya numerals, because very negligible work have been done and we wanted to explore the possibility of applying the two said features extraction techniques and compare the results of recognition for BPN and SVM.

IV. DATA SET COLLECTION AND PREPROCESSING

We have collected 20000 handwritten Odiya numerical from 200 different classes people. 100 numerical are collected from each person 10 of each digits starting from 0 to 9 in a pre-designed sheet. The collected documents are scanned using HP-scan jet 5400c at 300 dpi, which is usually a low noise and good quality image. The digitized images are stored as binary images in the bmp format and are then preprocessed.

Figure. 1 Sample datasheet of odiya handwritten numerals

The preprocessing stage involves conversion into binary, boundary extraction, noise reduction, framing and missing data recovery. The input numeral image is normalized after to size 30*30 for each handwritten numeral image.

A. Convert to binary:

The grey scale image is converted into binary image by representing the information in terms of 1s and background as 0s.

B. Boundary extraction:

The outer boundary of the image is obtained. The boundary of the binary image gives the outer shape of the numerals. The shape information are useful for the extraction of the features of the image in the chain code technique.

C. Noise reduction:

Noises are the data pattern which is not required and is present along with required numeral data pattern. A noise reduction function is called to rectify this problem. This noise is generally due to misprint or any undesirable mark present on the data sheet along with the required data or any noise inserted during scanning process.

D. Framing:

The binary image obtained after noise reduction consists of some additional bits information along with the required information of the image that doesn't add to the shape of the numerals. Processing the whole document not only consumes unnecessary time but also reduces the precision of the processed output. Those bits information are required to be removed, thus framing is done.

Basically, framing separates the character from the rest of the space. It is done by taking four bounds- left, right, upper, lower. The threshold value is assumed for each bound. The final output image of framing will contain the pixels carrying the information about the numeric characters and those enclosed within the character, eliminating the outer space.

E. Missing data recovery:

It might happen that the character may contain certain missing pixel information either due to decrease picture intensity or due to human error. These are recovered during the missing data recovery phase.

V. FEATURE EXTRACTION TECHNIQUES

For the classification to be more effective and efficient a well-defined feature extraction algorithm should be used. There are different algorithm available for feature extraction like zone based algorithm, template matching algorithms, fourier descriptors, geometric moment invariants, gradient features, freeman chain codes, etc,. We have used two well defined algorithms of feature extraction in our research work. They are:-

Fourier Descriptors

Normalized Chain Codes Methods

A. Fourier Descriptors:

Fourier descriptors are widely used in various applications for the shape discrimination. In this paper we deal about the recognition of the handwritten numerals by extracting the features from the shape of the binary image of the numerals. So we have used the algorithm which computes the fourier descriptors of the image. The coefficient of the fourier transform of the numeral image forms the fourier descriptors.

In fouirer transform, we converted the spatial domain of the numerals to its frequency domain. Fourier transform is computed along the coordinates of the pixel of the image which gives the information about its shape. The inputs for the computation of the transform are the coordinates of the pixel which gives information of the shape.

If we consider the kth pixel of the image, then its coordinate is given by (xk, yk). We consider the (column, row) coordinates of the point not as cartesian coordinates but as those in complex plane by writing

$f(k) = x(k) + i\, y(k)$

to get the fourier descriptor of this point we compute its discrete fourier transform .

$F(u)=1/K \sum f(k) \exp(-j2\pi uk/K) \qquad u = 0, 1, \ldots, K-1$

The coefficients of F(u) are the fourier descriptors which represents the shape in frequency domain. According to [4], the lower frequency descriptors contain information about the general features of the shape, and the higher frequency descriptors contain information about finer details of the shape. The high frequency information that describes the small details of the shape is not so helpful in shape discrimination, and therefore, they are ignored. In the process we obtain a large number of coefficients but a subset of it is enough to capture the shape of the numerals. These fourier descriptors have their characteristics as invariant to scale, translation and rotation. The coefficients thus obtained forms the feature vectors which are further used for the classification and recognition.

Algorithm for feature extraction using Fourier descriptors

Input: Gray scale handwritten numeral image.

Output: Finding 32 features for classification and recognition.

B. Method Begins:

a. The gray scale image is converted to binary image with numerals representing binary 1 for information and background 0.
b. The noise from the binary image is removed.
c. To extract the information of the shape of the numeral characters thinning is done to the binary image to bring a boundary.
d. To bring uniformity among the numerals a bounding box (frame) is fitted to the numeral image to a size of 30*30 pixels.
e. The pixels are represented in the complex plane, where the column co-ordinate is the real part and the row co-ordinate is the imaginary part.
f. The fourier transform is computed and the invariant fourier descriptors are obtained. The invariance is obtained by -nullifying the 0-th fourier descriptor (position invariance), dividing all fourier descriptors by the magnitude of the 1-st fourier descriptor (size invariance) and only considering the magnitude of the fourier descriptors (orientation and starting point

invariance). By applying this normalization, the 0-th fourier descriptors do not provide any information. Hence these are eliminated [4].

C. Method Ends:

a. Normalized Chain Code:

Chain codes are one of the techniques which deal with the algorithm that represents shape of the image by forming a boundary. The boundary formed by this method is a connected sequence of straight line segments of specified length and direction. The information about the shape of the image is obtained by encoding each connected components along the boundary separately. Each direction forming the boundary has a predefined specified value. For each such region, a point on the boundary is selected and its coordinates are transmitted. Any point in the image can be considered as the starting position and the value for each direction from the starting position is assigned. The encoder moves along the boundary of the region and, at each step, transmits a symbol representing the direction of this movement [8]. This process will continue until the encoder returns to the position from where it has started.

The process of connecting the pixels by finding the next component can be done by clockwise or anti clockwise direction along the extracted boundary of the image. The code is assigned to the new pixel depending upon the value of location of its previous location. The connection can be 4-connected or 8-connected. In this paper for our experiment, we have considered 8-connected chain codes and the boundary are formed in clock wise direction.

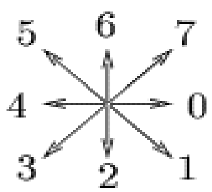

Figure 2. 8-connected chain code in clockwise direction.

According to [4], it is to be noted that different length of chain codes are obtained depending upon the size of the handwritten numerals. So in order to bring the uniformity we need to normalize the chain code obtained.

The frequency of the chain codes of the numerals are first found and represented as the frequency vector. Frequency vector contains the number of times of occurrence of the codes 0, 1, 2 … 7 in the boundary of the numerals. Then the frequency is normalized to get the normalized vectors. By concatenating the frequency vectors with the normalized vectors we get the feature vectors. Thus we extract 16 features vectors which are then classified for the recognition and are then experimented.

For example, let us consider A1 containing the chain codes that are obtained by traversing through the boundary of the binary image of the numeral in clockwise direction with 8-connected.

A1= [1 2 2 3 5 5 5 7 7 6 6 0 0 0 1 1 1 1 3 3 3 4 4 4 7 2 2 1 1 2 2 4 4]

The vector A1 contains huge dimensions. To reduce the dimensions, the frequency of occurrence of the chain code is calculated and represented by vector A2.

A2= [3 7 6 4 5 3 2 2]

The frequency vector is normalized by the following formula:

A3 = A2 / | A1|, where |A1|= ∑ A2

So, here we get A3 = [0.093 0.218 0.187 0.125 0.156 0.093 0.062 0.062]

Now in order to obtain the feature vector A2 and A3 are concatenated. Thus we get 16 feature vectors. These 16 features will be fed to the BPN and SVM for classification and reorganization.

Algorithm for feature extraction using chain code

Input: Gray scale handwritten numeral image

Output: Normalized chain code of length 16.

D. Method Begins:

a. The gray scale image is converted to binary image with numerals representing binary 1 and background 0.
b. The noise from the binary image is removed
c. To get the shape information the boundary of the image is extracted.
d. To bring uniformity among the numerals a bounding box (frame) is fitted to the numeral image to a size of 30*30 pixels.
e. The boundary that is extracted is connected clockwise assuming a starting pixels and connecting to the next pixel taking the previous location into account to generate a sequence of 8 dimensional chain codes.
f. The chain codes thus obtained are large in number, so it is normalized and thus the feature vectors are obtained.

E. Method Ends:

VI. EXPERIMENTAL RESULTS AND COMPARISON

Experiments are carried out over the feature vectors of 20000 handwritten numeral images. The feature vectors are obtained by the fourier descriptors method and normalized chain code method as described in the method 1 and method 2 of the section 5 of the paper. The feature vectors obtained from each method are then processed individually for classification to get the corresponding recognition rate and the results are then compared.

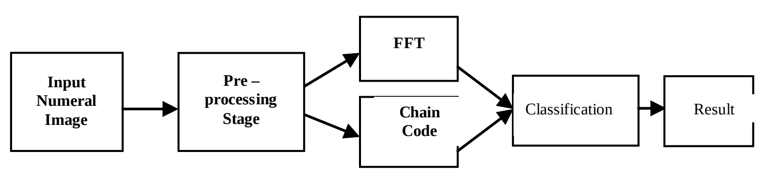

Figure 3 Overview of the recognition and classification process

RESULT 1. FOURIER DESCRIPTORS

Table-1 Recognition Rate for 20000 Data using BPN

NUMBER OF TRAINING SAMPLES	NO. OF HIDDEN LAYER NODES	EPOCH	RECOGNITION RATE (in %)
20000	10	166	73.33
20000	20	224	80.00
20000	30	149	77.33
20000	**40**	**289**	**84.33**
20000	50	233	83.33

Table-2 Recognition Rate for 20000 Data using SVM

ODIYA NUMERALS	RECOGNITION RATE (in %)
0	97.00
1	91.00
2	92.00
3	94.50
4	89.50
5	90.00
6	92.00
7	89.50
8	98.50
9	100
Average Recognition Rate	**93.4**

RESULT 2: NORMALIZED CHAIN CODE

Table-3 Recognition Rate for 20000 Data using BPN

NUMBER OF TRAINING SAMPLES	NO. OF HIDDEN LAYER NODES	EPOCH	RECOGNITION RATE (in %)
20000	10	122	73.97
20000	20	161	62.32
20000	**30**	**131**	**83.63**
20000	40	186	73.33
20000	50	165	79.33

Table-4 Recognition Rate for 20000 Data using SVM

ODIYA NUMERALS	RECOGNITION RATE (in %)
0	94.50
1	94.00
2	91.50
3	94.00
4	92.00
5	95.25
6	94.00
7	89.50
8	97.50
9	93.50
Average Recognition Rate	**93.57**

COMPARISON

Table: 5

FEATURE EXTRACTION TECHNIQUES	BPN	SVM
Fourier Descriptors	84.33%	**93.4%**
Normalized Chain Code	83.63%	**93.57%**

VII. CONCLUSION

In this paper we have implemented two feature extraction methods- fourier descriptors and normalized chain code for the recognition of Odiya handwritten numeral recognition. We have used back propagation neural network and support vector machines for classification. Out of all the features extracted we have taken 15440 training samples and 4560 testing samples for our experimental analysis. The classification is done separately for the two methods and the recognition rate are compared. It is observed that good recognition rate is achieved through the support vector machines classifier for both the feature extraction methods.

VIII. REFERENCE

[1]. Simon Haykin, Neural Network (a comprehensive foundation), 2nd edition, vol. 2, Prentice Hall International , 2000

[2]. S.V. Rajashekararadhya and Dr.P.Vanaja Ranjan, “Handwritten Numeral/Mixed numerals recognition of South-Indian scripts: the zone based feature extraction method”, Journal of Theoretical and Applied Information Technology, vol. 7, No 1, pp 063-079, 2009.

[3]. S.V. Rajashekararadhya and Dr.P.Vanaja Ranjan, “Efficient Zone based feature extraction algorithm for handwritten numeral recognition of four popular south Indian scripts”, Journal of Theoretical and Applied Information Technology, vol.7, 2008.

[4]. G. G. Rajput and S.M. Mali , “ Marathi Handwritten Numeral Recognition using Fourier Descriptors and Normalized Chain Code” IJCA Special Issue on “Recent Trends in Image Processing and Pattern Recognition” , 2010

[5]. Luiz S. Oliveira and Robert Sabourin, “Support Vector Machines for handwritten numerical string recognition”, supported by The National Council for Scientific and Technological Development , Pontificia Univ. Catolica do Parana, Curitiba, Brazil, 26-29 oct ,2004, pp 39-44

[6]. Shailendra Kumar Srivastava and Sanjay S. Gharde , “Support Vector machine for handwritten Devanagiri numeral recognition “, International journal of computer applications, volume 7,2010.

[7]. U. Pal, S. Kundu, Y. Ali, H. Islam and N. Tripathy, “Recognition of Unconstrained Malayalam handwritten numeral”, Indian Conference on Computer Vision , Graphics and Image Processing, Indian Statistical Institute, Kolkata ,16-18 Dec, 2004.

[8]. Nor Amizam jusoh and Jasni Mohammad Zain, “Application of Freeman Chain Codes: An alternative recognition techniques for Malaysian car plates”, International journal of computer science and network security, Vol.9, no.11, 2009.