Handwritten Gurmukhi Numeral Recognition using Zone-based Hybrid Feature Extraction Techniques

Gita Sinha Rajneesh Rani Renu Dhir

Department of Computer Science and Engineering
Dr B R Ambedkar National Institute of Technology
Jalandhar- 144011, Punjab (India)

ABSTRACT
This paper presents an overview of Feature Extraction techniques for off-line recognition of isolated Gurumukhi numerals/characters. Selection of Feature Extraction method is probably the single most important factor in achieving high performance in pattern recognition. Our paper presents Zone based hybrid approach which is the combination of image centroid zone and zone centroid zone of numeral/character image. In image centroid zone character is divided into n equal zone and then image centroid and the average distance from character centroid to each zones/grid/boxes present in image is calculated. Similarly, in zone centroid zone character image is divided into n equal zones and centroid of each zones/boxes/grid and average distance from zone centroid to each pixel present in block/zone/grid is calculated. SVM for subsequent classifier and recognition purpose. Obtaining 99.73% recognition accuracy.

General Terms
Feature extraction method, Image centroid zone, zone centroid zone, support vector classifier (SVM)

Keywords
Gurmukhi Script, Image processing, Pattern recognition, SVM.

1. INTRODUCTION
Optical Character Recognition (OCR) is Electronic conversion of Handwritten/Typewritten or printed image in machine encoded text. Pattern Recognition systems are used in many fields that have difference in nature. Handwritten digit, character and word recognition were introduced into this domain. Most researches have been done in Latin languages. OCR consists of many phases such as Pre-processing, Segmentation, Feature Extraction, Classifications and Recognition. The input of one step is the output of next step. Pre-processing consist of these operations slant correction, normalization and thicking and these have been adopted for the purpose of Feature Extraction the zone based method is used [2]. The last phase Classification SVM (Support Vector Machine) have been used as a classifier . The flow chart of a typical OCR can be shown as Figure 1.

Recognition of handwritten characters/numeral is one of the most interesting topics in pattern recognition. Applications, of OCR is in different area especially digit recognition, which deals with postal mail sorting, bank check processing, form data entry, vehicle plate recognition, postal address block detection and recognition ,camera OCR etc. For these applications, the performance (accuracy and speed) of digit recognition is most important factor. While in pattern classification and machine learning communities, the problem of handwritten digit recognition is a good example to test the classification performance [3]. HCR is very valuable in terms of the variety of applications and also as an academically challenging problem. When HCR is used as a solution for inputting regional language data and also as a solution for converting paper information to soft form, the internet can be enriched with regional information, so that the digital divide can be minimized. It also facilitate solution to processing large volumes of data automatically, for example, in processing hand-filled application forms into machine printed character /number . Hence many research work are going on different scripts . But on Indian language scripts, very scanty literatures are available [4].

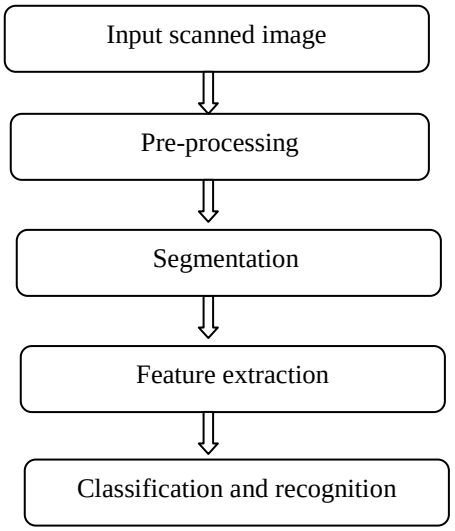


Figure 1: Phases in OCR

We have studied various research paper which reveals that the difficulty of the problem has two aspects. The first is attributed to the writer variations in style, size, shape, ink color, ink flow and thickness, digitization imperfections etc. The second is the deficiencies of the particular method used for feature extraction. Indian scripts share a large number of structural features due to common Brahmi origin. The differences between the scripts are primarily on their form of writing. The written form has more curves than straight or slant lines and has many similarities among different character/numeral of the same scripts and also between the scripts of different languages. The scripts of some languages like Hindi, Marathi, Punjabi etc, are same and there is many similarities between Kannada and Telugu. We used Gurmukhi script for our experiments. Gurmukhi characters are curved in nature with some kind of symmetric structure observed in the

shape. This information can be best extracted as a feature if we extract statistical features from the images.

As for the problem of OCR in low quality image different approach have been developed. The first one is to binarize the gray scale image by choosing the appropriate threshold and then perform feature extraction on binary image. However binarization process need details information of image degradation to distinguish the character stroke pixel and the background pixel in the image .For low quality image it is imposable to get such an degradation model so the binarization process will definetly result in loss of information , broken stroke and noise into binarized image. Another approaches directly applied on gray scale image without Binarization. This will avoid above disadvantage from binarization step[5].Mainly, character recognition machine will takes the raw data that for further implements. On the basis of data acquisition process, character recognition system can be classified into following categories.

1. Online Character Recognition
2. Offline Character Recognition

Off-line handwriting recognition refers to the process of recognizing words that have been scanned from a surface (such as a sheet of paper) and are stored digitally in gray scale format. After being stored, it is conventional to perform further processing to allow superior recognition.

In Online handwritten character recognition, the handwriting is captured and stored in digital form via different means. Usually, a special pen is used in conjunction with an electronic surface. As the pen moves across the surface, the two- dimensional coordinates of successive points are represented as a function of time and are stored in order. It is generally accepted that the on-line method of recognizing handwritten text have achieved better accuracy than off-line Recognition. This may be attributed to the fact that more information may be captured in the on-line case such as the direction, speed and the order of strokes of the handwriting. The major difference between Online and Offline Character
Recognition is that Online Character Recognition has real time contextual information but in case of offline recognition of pattern is imaginary.

This difference generates a significant divergence in processing architectures and methods. The offline character recognition can be further grouped into two types

- Magnetic Character Recognition (MCR)
- Optical Character Recognition (OCR)

In MCR, the characters are printed with magnetic ink. The reading device can recognize the characters according to the unique magnetic field of each character. MCR is mostly used in banks for check authentication. OCR deals with the recognition of characters acquired by optical means, typically a scanner or a camera. The characters are in the form of pixelized images, and can be either printed or handwritten, of any size, shape, or orientation. The OCR can be subdivided into handwritten character recognition and printed character recognition. Handwritten Character Recognition is more difficult to implement than printed character recognition due to diverse human handwriting styles and customs. In printed character recognition, the images to be processed are in the forms of standard fonts like Times New Roman, Arial, Courier, etc.

In the literature survey we have found that numbers of authors have attempted to recognize the Handwritten gurmukhi Numerals usung different techniques.

In 2010 Rafael M. O. Cruz et all. [2] have used Two techniques, Modified Edge Maps and Multi Zoning, are proposed feature extraction algorithm for character recognition . MLP as a classifier .

In 2011 Mahesh Jangid have proposed these method for feature extraction Zonal density, Projection histogram ,Distance Profiles, Background Directional Distribution (BDD) available [8], and SVM for classification and he have got 98%,99.1%and 99.2% of accuracy.

In 2011 Poulami Das et all. [10] have used binary tree based classifier and multilayer perceptron method on bangla character recognition. Result tested on 4423 dataset.They have obtained overall 90.00% of accuracy.

In 2011 Sushama Shelke [11] This paper presents a novel approach for recognition of unconstrained handwritten Marathi compound characters. The recognition is carried out using multistage feature extraction and classification scheme. They have obtained 96.14% and 94.22% recognition accuracy

In 2010 Shailedra Kumar et. al.[12] using Support Vector Machine for Handwritten Devanagari Numeral Recognition. Moment Invariant and Affine moment Invariant techniques are used as feature extraction . This linear SVM produces 99.48% overall recognition rate which is the highest among all techniques applied on handwritten Devanagari numeral recognition system.

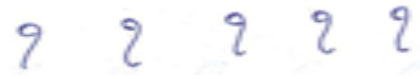
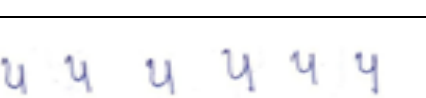
Pritpal Singh et.al.[13] used Feature Extraction and Classification Techniques in O.C.R. Systems for handwritten Gurmukhi Script – A Survey. They have used Zoning method for feature extraction and SVM, KNN for classification to obtained 72.54% recognition accuracy. Zoning density and background Directional distribution features extraction and SVM with RBF kernel which provide 95 % of accuracy.

Omid Rashnodi et all.[15] applied box approach method on Persian numeral. SVM classifier with linear kernel for recognition. Feature sets consists of 163 dimensions, which are the average angle and distance pixels which are equal to one in each box the box approach. the result has been evaluated on 6,000 sample of numeral. They was obtain 98.94% accuracy.

2. DATASET COLLECTION

We have used same dataset that has been used in this paper [14]. In which 1500 Gurmukhi numeral has been collected from 15 different witters. Document were scanned using HP Jet scanner .These all samples are stored in JPEG/BMP format. The Handwritten Gurmukhi numeral dataset are shown in Table-1.

Table1 Handwritten Gurmukhi numeral samples

0	
1	
2	
3	
4	
5	
6	
7	
8	
9	

3. PHASES IN OCR

OCR consist of following stage pre-processing, segmentation, feature extraction, classification and recognition.

3.1 Pre-processing

In preprocessing operations sample image are converted into gray scale .Then we have applied these techniques, Gray scale image are converted into binary image using threshold value obtained by Otsu's method, filtering operation, morphological operation, removal of noise having less than 30 pixels, Binarization, contour smoothing, skew detection, and skeletonization of a digital image so that subsequent algorithms along the road to final classification can be made simple and more accurate [14]. The corresponding objectives of Pre-processing methods are as follows: -

3.1.1 Normalization

The input numeral image is normalized to equal size.

3.1.2 Noise removal

The major objective of noise removal is to remove any unwanted bit-patterns, which do not have any significance in the output.

3.1.3 Contour Smoothing

The objective of contour smoothing is to smooth contours of broken and/or noisy input characters

3.1.4 Skew detection

Skew detection is one the first operations to be applied to scanned documents when converting data to a digital format. Its aim is to align an image before processing because text segmentation and recognition methods require properly aligned next lines.

3.1.5 Skeletonization

Skeletonization is the process which reduces the width of a line . This process can remove irregularities in Character. This methods only applied on character stroke which have only one pixel width. It requires less memory to store information about input numeral. This method also requires less processing time.

After pre-processing phase, a cleaned image is available that goes to the segmentation phase.

The raw data, depending on the data acquisition type, is subjected to a number of preliminary processing steps to make it usable in the descriptive stages of character analysis. Pre-processing aims to produce data that are easy for the HGCR system to operate accurately [6].

4. SEGMENTATION

After the pre-processing steps there is a very important and difficult task of OCR that is Script Segmentation. Script segmentation is done by executing the following operations: Line segmentation, Word segmentation character segmentation[7]

5. FEATURE EXTRACTION-

The selection of good feature set is the most the important aspect of handwritten pattern recognition. This method provide the ease of implementation and good recognition. Used the following sets of extracted features to recognize Gurmukhi numerals. In the next section define these algorithm step-by-step .The following paragraph explained the details about feature extraction method.

Compute the centroid of image (numeral/character). This image is further divided into 100×100 equal zones where size of each zone is (10×10). Then compute the average distance from image centroid to each pixel present in the zones/block. Obtaining 100 feature vector of each image. Similarly in ZCZ divide image into n equal zones and calculate centroid of each zones. Then compute the average distance from the zone centroid to each pixel present in zones. There could be some zones that are empty then the value of that particular zone is assumed to be zero. Repeat these procedure for all zones present in image(numeral/character) [2].

This paper present efficient zone based feature extraction algorithm for handwritten numeral recognition of four popular south Indian scripts that has been define in this paper [2]. But we have apply the same method on Gurmukhi scripts. Algorithm 1 provides Image centroid zone (ICZ) based distance metric feature extraction system and Algorithm 2 provides Zone Centroid Zone (ZCZ) based Distance metric feature extraction system. Algorithm 3 provides the combination of both (ICZ+ZCZ) feature extraction system. The following are the algorithms to show the working procedure of our feature extraction methods and also in fig-2.

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

Input: Image(character/numeral) are Pre-processed.

Output: Extract the Features for Classification and Recognition.

Algorithm

Step 1: Calculate centroid of input image.

Step 2: Division of input image in to **n** 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: Computation of average distance between these points.

Step 6: Repeat this procedure sequentially for the entire zone present in the image.

Step 7: Obtaining **n** such feature for Classification and recognition process.

Ends.

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

Method Begins

Step 1: Division of input image in to **n** equal zones.

Step 2: Compute centroid of each zones.

Step 3: Compute the distance between the zone centroid to each pixel present in the zone.

Step 4: Repeat step 3 for the entire pixel present in the zone/box/grid.

Step 5: computation of average distance is between these points present in image.

Step 6: This procedure are sequentially Repeat for the entire zone.

Step 7: obtaining **n** such features for classification and recognition.

Ends.

Proposed hybrid Algorithm 3: Which is the combination of both **(**ICZ + ZCZ) based Distance metric feature extraction system.

Input: Image(numeral/character) are Pre-processed.

Output: Extracted Features for Classification and Recognition

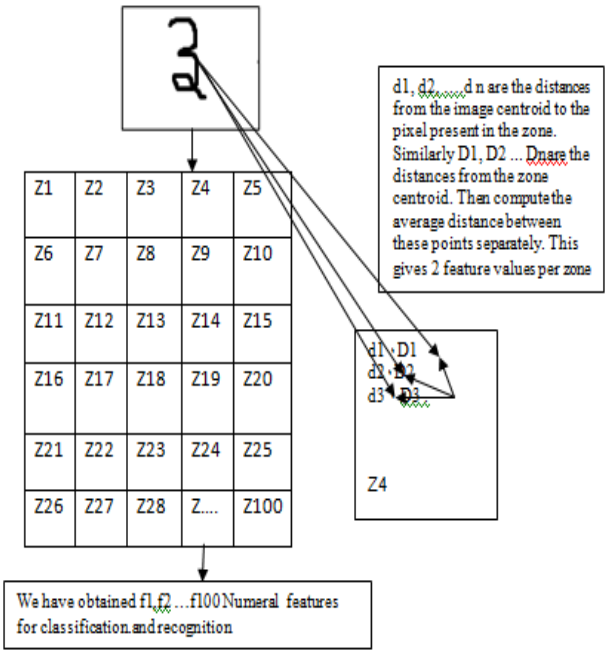


Figure 2 All procedure to extract feature from numeral image.

Method Begins

Step 1: Compute Centroid of input image.

Step 2: Division of input image into **n** equal zones.

Step 3: Compute the distance between th image Centroid to each pixel present in the zone.

Step 4: Repeat step 3 for the entire pixel present in the zone.

Step 5: Computation of average distance between these all points in the image.

Step 6: Compute centroid of the zone/block .

Step 7: Computation of the distance between the zone centroid to each pixel present in the zone.

Step 8: Repeat step 7 for the entire pixel present in the zone.

Step 9: Computation of average distance between these points.

Step 10: Repeat the steps 3-9 sequentially for the entire zone.

Step 11: Obtaining **2*n** such feature for classification and recognition.

Ends.

6. CLASSIFICATION AND RECOGNITION

We have used Support Vector Machines (SVM) 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 have capability of learning is to achieve good generalization performance. Which is objective of any machine, given a finite amount of training data by 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, for each given input, which of two possible classes forms the 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 hyper-plane as the decision plane, which separates the positive and negative patterns with the largest margin. The process of rearranging the objects is known as mapping (transformation). Rearranging the object, using a set of mathematical functions, known as kernels. There are some common Kernel functions that include the linear kernel, the polynomial kernel and the radial basis function (RBF) and sigmoid [9]. We have obtained such multiclass SVM tool LIBSVM available at [1]. We have used RBF (Radial Basis Function) kernel which is also common choice, in our recognition. RBF has single kernel parameter gamma (g or γ). Additionally there is another parameter with SVM classifier called soft margin or penalty parameter (C).

SVM have proved to achieve good generalization performance by the use of concept of basis, without knowledge of the prior data [15].

Table 2. Gurmukhi numeral accuracy using five fold cross validation .

S .No.	Parameters	Accuracy
1.	C=1 γ =.001	98.86%
2.	C=2 γ =.004	99.40%
3.	C=8 γ=.008	99.53%
4.	C=4 γ=8	99.60%
5.	C=4 γ=16	99.73%

7. EXPERIMENTAL RESULT

We have applied above Recognition strategy on Gurmukhi numeral. The dataset consist of 1500 samples of Gurmukhi numeral .There are 150 Sample for each of the numeral . we have used Zone based Feature Extraction techniques on this samples and practiced on different size of image. We obtained different recognition accuracy. The highest accuracy obtained from image size 100×100 and zone (10×10). We have obtained 200 Feature Vector from both of the methods. The accuracy depends upon size of image and number of Feature vector we have obtained from the image. The numeral recognition accuracy also depend on SVM parameter(C, γ) .We have obtained 99.73% of accuracy on Gurmukhi numeral with the value of C and γ.

8. COMPARISON WITH EARLIER APPROACHES

We compared the result using Projection histogram Distance profile with Image centroid zone and Zone centroid zone on the same dataset. Table 3 show the experimental result

Table 3 Previous results on handwritten Gurmukhi numerals

Proposed by	Feature extraction techniques	Classification techniques	RR
Kartar Singh Sidharth [13]	Projection histogram	SVM with RBF kernels	99.20%
Kartar Singh Sidharth [13]	Distance profile	SVM with RBF kernels	99.13%
Our work	Zone based	SVM with RBF kernels	**99.73%**

9. CONCLUSION

In this paper we have used a Zone Based Hybrid Feature Extraction Techniques which is the combination of Image Centroid Zone and Zone Centroid Zone by S.V. Rajashekararadhya and Dr P. Vanaja Ranja .Our Experimental Results proved that ZCZ method provide better recognition accuracy than ICZ.

10. ACKNOWLEDGMENT

We are very thankful to Kartar Singh Siddharth for his sincere help towards the provision of collected Gurmukhi Numeral dataset for our experiment.

11. REFERENCES

[1] Chih-Chung Chang and Chih-Jen Lin LIBSVM : A Library of Support Vector Machine software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[2] Rafael M. O. Cruz, George D. C. Cavalcanti and Tsang Ing Ren " An Ensemble Classifier For Offline Cursive Character Recognition Using Multiple Feature Extraction Techniques" *978-1-4244-8126-2/10/$26.00 ©2010 IEEE*

[3] L R Ragha, M Sasikumar "using moment feature for gabor directional image for kannada handwritten character recognition" International Conference and Workshop on Emerging Trends in Technology (ICWET 2010) – TCET, Mumbai, India

[4] Xuewen Wang Xiaoqing Ding ,Changsong Liu "Gabor filters- based feature extraction on character recognition" pattern recognition 38 (2005) 369-379.

[5] Naveen Garg "Handwritten Gurmukhi Numeral Recognition using Nueral Network " M.tech Thesis, Thapar University, Patiala

[6] Arif Billah Al-Mahmud Abdullah and Mumit Khan " a survey on script segmentation for bangla ocr" Working Papers 2004-2007.

[7] Mahesh Jangid Kartar Singh, Renu Dhir Rajneesh Rani "Performance Comparison on Devanagari Handwritten Numeral Recognition" International Journel of Computer Application (0975-8887) volume-22 No.-1, May 2011 .

[8] Statsoft electronic statistic textbook creator of statistica data analysis and service http://www.statsoft.com/ textbook/support-vector machines/

[9] Poulami Das Suchandra Paul Ranjit Ghoshal "Recognition of Bangla Basic Characters using Multiple Classifiers"Handwritten Numeral/Mixed Numerals Recognition of South Indian: Zona based Feature Extraction Method", *978-1-4577-1386-611$26.00©2011 IEEE*

[10] Sushama Shelke, Shaila Apte "A Multistage Handwritten Marathi Compound Character Recognition Scheme using Neural Networks and Wavelet Features" *International Journal of Signal Processing, Image Processing and Pattern Recognition Vol. 4, No. 1, March 2011*

[11] Shailedra Kumar Shrivastava, Sanjay S. Gharde " Support Vector Machine for Handwritten Devanagari Numeral Recognion " International Journel of Computer Application (0975-8887) Volume 7-No. 11,October 2010

[12] Pritpal singh*, sumit budhiraja, " Feature Extraction and Classification Techniques in O.C.R. Systems for handwritten Gurmukhi Script – A Survey " , International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 Vol. 1, Issue 4, pp. 1736-1739

[13] Kartar Singh Siddharth Renu Dhir Rajneesh Rani "Handwritten Gurmukhi Numeral Recognition using Different Feature Sets" International Journal of Computer Application (0975-8887) Vol. 28 No.-2 , August 2011

[14] H. Swethalakshmi, Anita Jayaraman, V. srinivasa Chakravarthy , C. Chandra Sekhar , "Online Handwritten Recognition of Devanagari and Telgu Character using Support Vector machine".

[15] Omid Rashnodi, Hedieh Sajedi , Mohammad Saniee "Using Box Approach in Persian Handwritten Digits Recognition" *International Journal of Computer Applications (0975 – 8887) Volume 32– No.3, October 2011*