Volume 2, Issue 6, June 2012 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Handwritten Gurmukhi Character Recognition Using K-NN and SVM Classifier

Gita Sinha, Mrs. Rajneesh Rani, Prof. Renu Dhir

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

ABSTRACT-In this paper we present an overview of Feature Extraction techniques for off-line recognition of isolated Gurumukhi characters recognition. Selection of Feature Extraction method is probably the single most important factor in achieving high performance in pattern recognition. Our paper presents Zone based 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 calculate. We have used SVM and K-NN for subsequent classifier and recognition purpose. We obtained 95.11% and 90.64% recognition accuracy SVM and K-NN respectively.

Keywords: Gurmukhi Script, Image processing, pattern recognition , SVM, ICZ, ZCZ.

1. INTRODUCTION

Nowadays, recognition systems are used in many fields that have different nature. The optical character recognition (OCR) was started from the recognition of machine printed digits and characters and then it was developed to the recognition of machine printed words. Gradually, 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 consists 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) and K- nearest neighbor (K-NN) have been used as a classifier. The flow chart of a typical OCR can be shown as Figure 1.

Recognition of handwritten character 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].

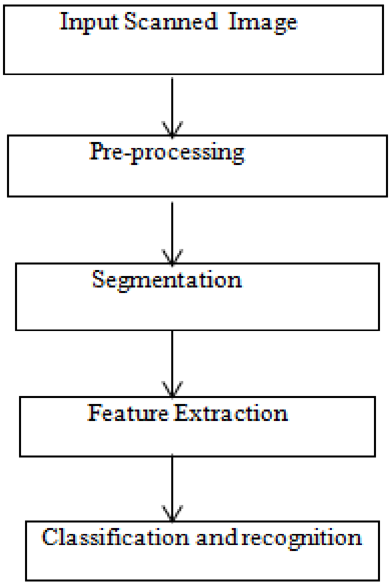


Figure 1: Phases in OCR

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 colour, 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 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.

In the literature survey we have found that numbers of authors have attempted to recognize the Handwritten Gurmukhi Character using different techniques.

M. K. Jindal [17] .In this paper, they have proposed a solution for segmenting horizontally overlapping lines. Whole document has been divided into strips and proposed algorithm has been applied for segmenting horizontally overlapping lines and associating small strips to their respective lines. The results reveal that the algorithm is almost 99% perfect when applied to the Gurmukhi script.

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 they have got 98%,99.1% and 99.2% of 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.

In 2009 S. V. Rajashekararadhya et.al.[10] have used the zone based feature extraction method on handwritten numeral/mixed numerals recognition of south-indian scripts. Feed forward back propagation neural network, and Support Vector Machine for classifier. They have obtained overall 98.9% of accuracy.

In 2008 S.V. Rajashekararadhya et. al [2] have used efficient zone based feature extraction algorithm for handwritten numeral recognition of four popular south indian scripts and Nearest neighbour ,Neural network as a classifier . They have got 98.5% of accuracy.

In 2007 M. Hanmandlu et.al [11] using Input Fuzzy Modeling for the Recognition of Handwritten Hindi Numerals. This paper presents the recognition of Handwritten Hindi Numerals based on the modified exponential membership function fitted to the fuzzy sets derived from normalized distance features obtained using the Box approach method . They have obtained 95% recognition accuracy.

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.

ZHAO Bin et. al [18] they have used Support Vector Machine for classification and its Application in Chinese check recognition system. The experiment on NIST numeral database and the actual check samples shows that comparison with other classifiers, SVM possesses better generalization ability. They have got 99.95% accuracy.

2. Gurmukhi Script

Gurmukhi, a derivative of Landa, is a type of script called an abugida. It was standardised by Guru Angad Dev in the sixteenth century and is designed to write the Punjabi language. This paper introduces the main concepts of the Gurmukhi script in relation to the Punjabi language. Gurmukhi has been adapted to write other languages (such as Sanskrit) but these adaptations will generally not be covered.

The Gurmukhi (or Punjabi) alphabet contains thirty-five distinct letters. These are shown in Figure 2. The first three letters are unique because they form the basis for vowels. Apart from Era, these characters are never used on their own. Gurmukhi follows similar concepts to other Brahmi scripts and as such, all consonants are followed by an inherent ‘a’ sound (unless at the end of a word when the ‘a’ is usually dropped) [18]. This inherent vowel sound can be changed by using dependent vowel signs which attach to a baring consonant. In some cases, dependent vowel signs cannot be used – at the beginning of a word or syllable for instance – and so an independent vowel character is used instead.

ੳ ਊੜਾ (ūṛā) u, ū, o	ਅ ਐੜਾ (aiṛā) a, ā, ai, au	ੲ ਈੜੀ (īṛī) i, ī, e	ਸ ਸੱਸਾ (sas'sā) sa [sə]	ਹ ਹਾਹਾ (hāhā) ha [hə]
ਕ ਕੱਕਾ (kakkā) ka [kə]	ਖ ਖੱਖਾ (khakhkhā) kha [kʰə]	ਗ ਗੱਗਾ (gaggā) ga [gə]	ਘ ਘੱਗਾ (ghaggā) gha [gə]	ਙ ਙੰਙਾ (ṅaṅṅā) ṅa [ŋə]
ਚ ਚੱਚਾ (caccā) ca [ʧə]	ਛ ਛੱਛਾ (chachchā) cha [ʧʰə]	ਜ ਜੱਜਾ (jajjā) ja [ʤə]	ਝ ਝੱਜਾ (jhajjā) jha [ʤə]	ਞ ਞੰਞਾ (ñaññā) ña [ɲə]
ਟ ਟੈਂਕਾ (taiṅkā) ṭa [ʈə]	ਠ ਠੱਠਾ (ṭhaṭhṭhā) ṭha [ʈʰə]	ਡ ਡੱਡਾ (ḍaḍḍā) ḍa [ɖə]	ਢ ਢੱਡਾ (ḍhaḍḍā) ḍha [ɖə]	ਣ ਣਾਣਾ (ṇāṇā) ṇa [ɳə]
ਤ ਤੱਤਾ (tattā) ta [tə]	ਥ ਥੱਥਾ (thaththā) tha [tʰə]	ਦ ਦੱਦਾ (daddā) da [də]	ਧ ਧੱਦਾ (dhaddā) dha [də]	ਨ ਨੱਨਾ (nannā) na [nə]
ਪ ਪੱਪਾ (pappā) pa [pə]	ਫ ਫੱਫਾ (phaphphā) pha [pʰə]	ਬ ਬੱਬਾ (babbā) ba [bə]	ਭ ਭੱਬਾ (bhabbā) bha [bə]	ਮ ਮੱਮਾ (mam'mā) ma [mə]
ਯ ਯੱਯਾ (yayyā) ya [jə]	ਰ ਰਾਰਾ (rārā) ra [rə]	ਲ ਲੱਲਾ (lallā) la [lə]	ਵ ਵੱਵਾ (vavvā) va [wə]	ੜ ੜਾੜਾ (rāṛā) ṛa [ɽə]
ਸ਼ ਸ਼ੱਸ਼ਾ (śaśśā) śa [ʃə]	ਖ਼ ਖ਼ੱਖ਼ਾ (k̲h̲ak̲h̲k̲h̲ā) k̲h̲a [χə]	ਗ਼ ਗ਼ੱਗ਼ਾ (ġaġġā) ġa [ɣə]		
ਜ਼ ਜ਼ੱਜ਼ਾ (zazzā) za [zə]	ਫ਼ ਫ਼ੱਫ਼ਾ (faffā) fa [fə]	ਲ਼ ਲ਼ੱਲ਼ਾ (l̲al̲l̲ā) ḷa [ɭə]		

Figure 2 Sample of Gurmukhi Constant Character.

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

3. 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 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].

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

4. 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. Script segmentation is done by executing the following operations: Line segmentation, Word segmentation and character segmentation [7]

5. FEATURE EXTRACTION-

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

We have computed 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 we have computed the average distance from image centroid to each pixel present in the zones/block. We have got 100 feature vector of each image. Similarly in ZCZ we have divided image into n equal zones and calculated 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. We repeat these procedure for all zones present in image(numeral/character) [2].

We have used 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 character recognition. 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 figure 3.

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

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 system.

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: Extract Features for Classification and Recognition.

Method Begins

Step 1: Compute Centroid of input image.

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

Step 3: Computation of the distance between the 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.

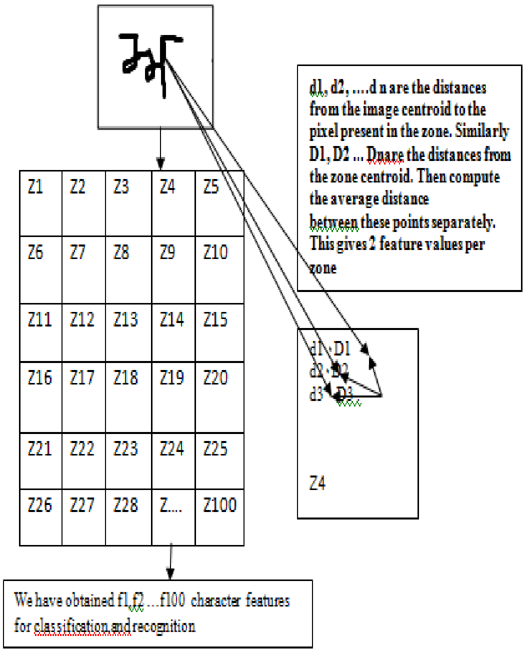
Step 8: Repeat step 7 for the entire pixel present in the zone.

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 features for classification and recognition.

Ends.



d1, d2, ...d n are the distances from the image centroid to the pixel present in the zone. Similarly D1, D2 ... Dnare the distances from the zone centroid. Then compute the average distance between these points separately. This gives 2 feature values per zone

We have obtained f1,f2 ...f100 character features for classification and recognition

Figure 3 All procedure to extract feature from gurmukhi character image "Era (a)".

6. CLASSIFICATION AND RECOGNITION

We have used Support Vector Machines (SVM) and K-nearest neighbor (K-NN) for the purpose of Classification and recognition.

A.SUPPORT VECTOR MACHINES

SVM is 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. The standard SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the input.

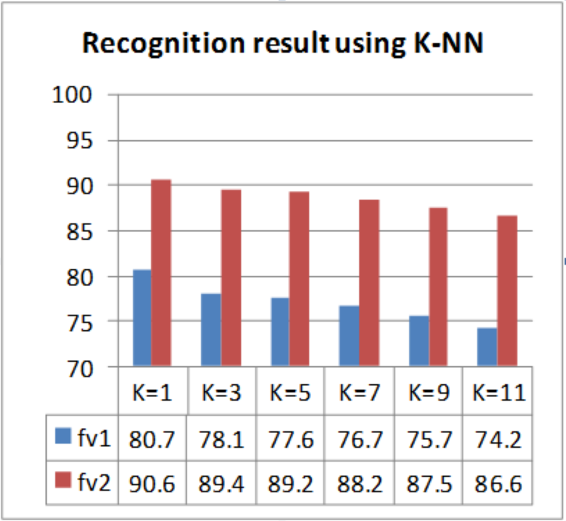
The process of rearranging the objects is known as mapping (transformation). Rearranging the object, using a set of mathematical functions, known as kernels. There are some common Kernel functions that include the linear kernel, the polynomial kernel, radial basis function (RBF) and sigmoid [9]. We have 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).

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

B. K- NEAREST NEIGHBOR (K-NN)

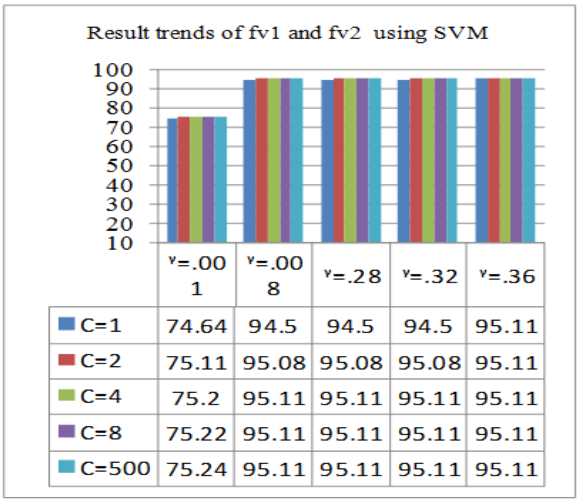
The k-nearest neighbor algorithm (k-NN) is a method for classifying objects based on closest training examples in the feature space. K-NN is a type of instance-based learning, or lazy learning where the function is only approximated locally and all computation is deferred until classification. The k-nearest neighbor algorithm is amongst the simplest of all machine learning algorithms: an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of its nearest neighbor available at [21].

In figure 4 that shows the best results are obtained with FV2 and second best results are obtained with FV1. Both of these feature sets are comprised of ICZ and ZCZ features respectively. These feature sets have 16 and 100 number of features respectively. Hence the increased recognition rate of FV2 is obtained at the cost of significantly increased features. The classification using K-NN is less time consuming because it is instant based learning methodology instead of eager learning used in SVM.



	K=1	K=3	K=5	K=7	K=9	K=11
🟦 fv1	80.7	78.1	77.6	76.7	75.7	74.2
🟥 fv2	90.6	89.4	89.2	88.2	87.5	86.6

figure 4. Recognition accuracy by K-NN



	ᵞ=.00 1	ᵞ=.00 8	ᵞ=.28	ᵞ=.32	ᵞ=.36
🟦 C=1	74.64	94.5	94.5	94.5	95.11
🟥 C=2	75.11	95.08	95.08	95.08	95.11
🟩 C=4	75.2	95.11	95.11	95.11	95.11
🟪 C=8	75.22	95.11	95.11	95.11	95.11
🟦 C=500	75.24	95.11	95.11	95.11	95.11

Figure 5. Gurmukhi character accuracy using five fold cross validation.

In figure 5 the significant difference is only at smaller values of γ (2 and 8) and even these variations are only observable at smaller values of C and at larger values of C (>2)_all the results are mostly coinciding for all range of γ. Most of the results of SVM listed are observed at larger range of C tested upto 500.It is clear from the table shown below the figure that highest result of our work is obtained for FV2 feature set. This highest result is 95.11% which is obtained at γ = 0.001 with combination of C = 4, 8,32, 64 and 500 (i.e. C ≥ 2).

EXPERIMENTAL RESULT

We have applied above Recognition strategy on Gurmukhi character. The dataset consist of 7000 samples of Gurmukhi character .There are 7000 Sample for each of the character. 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 recognition accuracy also depend on SVM parameter(C, γ) .We have obtained 95.11% and 90.64% of accuracy on Gurmukhi character with the value of **C >2 , γ=.1 , .01 and K=1,..11** by SVM and K-NN respectively.

Table-1 Comparison with earlier approach.

Proposed by	Features extraction Method	Classifier	Accuracy (dataset) (year)
Puneet Jhaji et all	Zoning	SVM	73.83% (2050) 2010
Ubeeka Jain et	Profiles, width, height, aspect ratio,	Neocognitron Neural Network	92.78% (15000) 2010
Anuj Sharma et al.	Strokes recognition and matching.	Elastic matching	90.08% (2008)
Munish Kumar et all.	Diagonal & transition feature	K-NN	94.12% (3500) 2011
Kartar Singh Siddharth et al	zonal density, projection histograms, distance profiles, BDD	K-NN, SVM, PNN	95.01% (7000) 2011
Our Work	Zonal based feature	SVM	95.11% (7000) 2011

CONCLUSION

In this paper we have used a Zone Based Feature Extraction Techniques which is the combination of Image Centroid Zone and Zone Centroid Zone by S.V. Rajashekararadhya and Dr P. Vanaja Ranja . This algorithm has shown a notable improvement for recognizing of handwriten Gurmukhi character. Our Experimental Results proved that ZCZ method provide better recognition accuracy than ICZ.

ACKNOWLEDGMENT

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

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]. S.V. Rajashekararadhya, Dr P. Vanaja Ranjan, . 2008 "efficient zone based feature extraction algorithm for handwritten numeral recognition of four popular south indian" journal of theoretical and applied information technology

[3]. Omid Rashnodi, , Hadieh Sajedi, Mohammad Saniee "using box box approach in persiam handwritten digit recognition" International Journal of Computer Applications (0975 – 8887) Volume 32– No.3, October 2011.

[4]. 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

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

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

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

[8]. 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 .

[9]. http://www.statsoft.com/textbook/support-vector machines/

[10]. V. Rajashekararadhya, P. Vanaja ranjan, "Handwritten Numeral/Mixed Numerals Recognition of South Indian: Zona based Feature Extraction Method" , 2005 - 2009 JATIT.

[11]. M. Hanmandlu, J. Grover, V. K.. Madasu, S. Vasikarla " Input Fuzzy Modeling for the Recognition of Handwritten Hindi Numerals "

International Conference on Information Technology (ITNG'07) 0-7695-2776-0/07 ,2007 IEEE.

[12].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

[13]. 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

[14].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

[15]. H. Swethalakshmi, Anita Jayaraman, V. srinivasa Chakravarthy , C. Chandra Sekhar , “Online Handwritten Recognition of Devanagari and Telgu Character using Support Vector machin”.

[16].Guca.sourceforge.net/.../introductiontogurmukhi.

[17].M. K. Jindal , R. K.Sharma, G. S. Lehal Segmentation of Horizontall Overlapping lines in Printed Gurmukhi Script” 1-4244-0716-8/06/ ©2006 IEEE.

[18].ZHAO Bin, LIU Yong and XIA Shao- Wei “Support

[19].Vector Machine and its Application in Handwritten

[20].Numeral Recognition” 0-7695-0750-6/00 2000 IEEE

[21].http:saravananthirumuruganathan.wordpress.com2010/05/17/a-detailed-introduction-to-k-nearest-neighbor-knn-algorithm/