# Design and Simulation of Handwritten Gurumukhi and Devanagri Numerals Recognition

Naveed Anjum
Dr, B.R Ambedkar
NIT Jalandhar

Tarun Bali
Dr, B.R Ambedkar
NIT Jalandhar

Balwinder Raj,Ph.D
Dr, B.R Ambedkar
NIT Jalandhar

## ABSTRACT

The work presented in this paper focuses on recognition of isolated handwritten numerals in Devanagari and Gurumukhi script. The proposed work uses four feature extraction methods like Zoning density, Projection histograms, Distance profiles and Background Directional Distribution(BDD). On the basis of these four types of features we have formed 10 feature vectors using different combinations of four basic features. This work uses Support Vector machines(SVM) for the classification of numerals. A total of 2000 samples of numerals are taken for Gurumukhi and Devanagari and we have attain a maximum recognition accuracy of 99.6% in case of Gurumukhi Numeral recognition and 99% for Devanagri Numeral recognition. In addition to SVM classifier , we have also used two similarity based classifiers Euclidean distance and Square chord distance for the classification purpose. With Euclidean distance ,a recognition accuracy of 99% and 91.67% is obtained for Gurumukhi and Devanagri numarals respectively. Similarly with Square Chord distance accuracy of 95.33% and 81.67% is obtained for Gurumukhi and devanagri numerals respectively

## Keywords

Character recognition, Feature extraction, support vector machine, classification.

## 1. INTRODUCTION

Optical Character Recognition(OCR) is the process of converting the scanned images of handwritten, typewritten or printed text into machine or computer editable text. [3]. The character recognition is classified into two main categories: Online line recognition and Offline recognition .On-line character recognition deals with a data stream which comes from a transducer while the user is writing. While the Off-line character recognition is performed after the writing is finished. The offline[2] character recognition is further classified as Printed and Handwritten. Earlier OCR was widely used to recognize printed or typewritten documents. But recently, there is an increasing trend to recognize handwritten documents. The recognition of handwritten documents is more complicated in comparison to recognition of printed documents It is because handwritten documents contains unconstrained variations of written styles by different writers even different writing styles of same writer on different times and moods. The recognition of handwritten numerals plays an important role in OCR research and development due to many potential applications, such as bank check processing, postal mail sorting,[22] automatic reading of tax forms and various handwritten forms. .Handwritten numeral recognition is an exigent task due to the restricted shape variation, different script style & different kind of noise that breaks the strokes in number or changes their topology However, as handwritten characters have a variety of styles, there is still room for researchers to develop new methods to increase recognition accuracy. Hand written OCR systems consist of five major stages as shown in figure 1 :

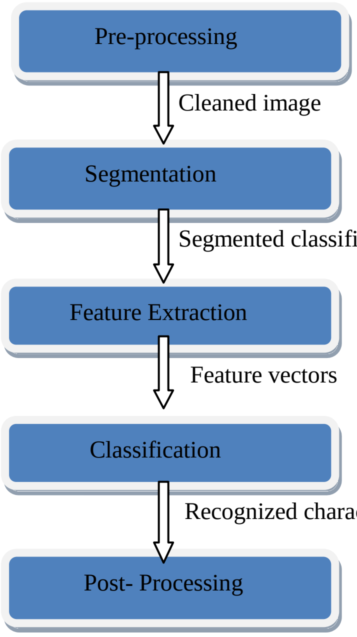

**Figure 1: Flow of OCR system**

## 1.1 Preprocessing

The pre-processing phase normally includes many techniques applied for binarization, noise removal, skew detection, slant correction, normalization, contour making and skeletonization like processes to make character image easy to extract relevant features and efficient recognition[4].

## 1.2 Segmentation

Segmentation partitions the digital image into multiple segments. It is used to decompose an image of a sequence of characters into sub images of individual symbols by segmenting words and lines [6].

## 1.3 Feature Extraction

Feature extraction is extracting information from raw data which is most [7] relevant for classification purpose. In feature extraction stage every character is assigned a feature vector to identify it. This vector is used to distinguish the character from other characters

## 1.4 Classification

Classification is the main decision making stage of OCR system. It uses the features extracted in the previous stage to identify the characters.[3] Classifiers are first trained by a training set of pattern samples to prepare a model which is later used to recognize the test samples. The training data should consist of wide varieties of samples to recognize all possible samples during testing. Some examples of generally practiced classifiers are- Support Vector Machine (SVM), K-Nearest Neighbor (K-NN), Probabilistic Neural Network (PNN).[8]

## 1.5 Post processing

The output of classification may contain some recognition errors. Post-processing methods remove these errors by making use of mostly two methods namely, dictionary lookup and statistical approach

## 2. DATA COLLECTION AND PREPROCESSING

A total of 2000 Gurumukhi numeral samples from 20 different writers were collected. Writers were provided with a plain A4 sheet and each writer was asked to write Gurumukhi numerals from 0-9 ten times . Similarly, the dataset size of 2000 devanagri numerals is used. The digitized images are stored as binary images in the JPEG format

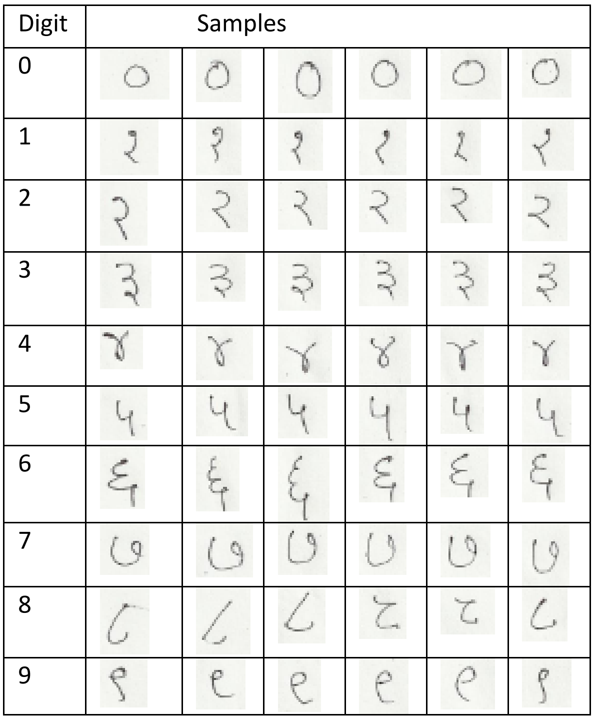
Digit
Samples
0
1
2
3
4
5
6
7
8
9

**Fig 2: Handwritten Gurumukhi scanned numerals samples**

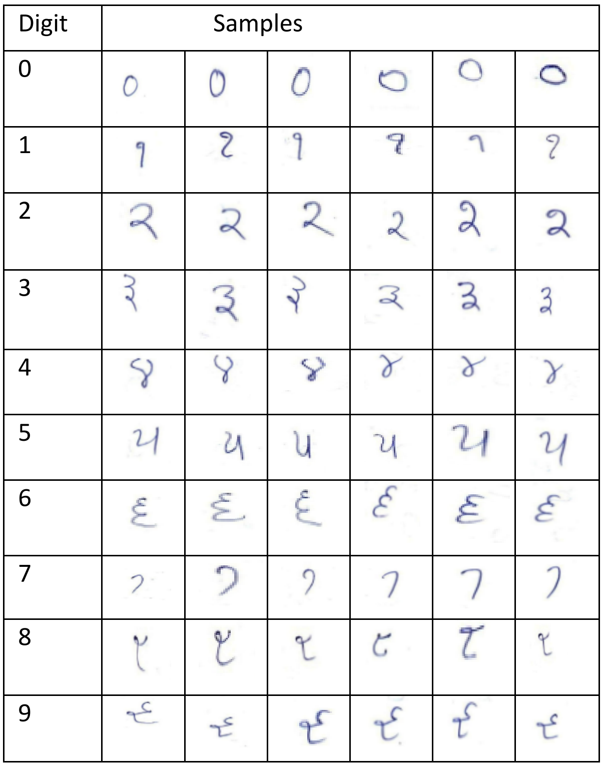
Digit
Samples
0
1
2
3
4
5
6
7
8
9

**Fig 3: Handwritten Devanagri scanned numerals samples**

Image Pre- processing is a process in which the scanned images of handwritten numerals are first converted to binary image and then various types of techniques are applied in order to remove noise in order to make the images ready for feature extraction and classification purposes. Pre-processing involves a series of operations that are being performed on the scanned images. These are discussed as follows:

## 2.1 Binarization

The scanned images of the numerals may be colour image format. So it is required to first convert it into gray level image before converting to binary image. Gray level image is one in which each pixel of the image is represented by intensity values lying between 0 and 1. But in binary images there are only twolevels 0 and 1. Black pixel is represented by 0 and white with 1.

## 2.2 Noise Removing

Scanning process may introduce noise that [6] may be in the form of disconnected line segments, blurred images etc. In order to remove noise special filters are used .

## 2.3 Skew correction

Deviation of the baseline of the text from horizontal direction is called skew. Document [19] skew often occurs during document scanning or copying. This effect visually appears as a slope of the text lines with respect to the x-axis. Skew lines are made horizontal and making proper correction in the raw image.

## 2.4 Slant correction

The character inclination that is normally found in cursive writing is called slant. Slant correction an important step in the pre-processing stage of handwritten character recognition. To correct the slant presented first we need to estimate the slant angle , then horizontal shear transform is applied to all the pixels of images of the character in order to shift them to the left or to the right.

## 2.5 Normalization

Normalization is required as the size of the numeral varies from person to person and even with the same person from time to time. The input numeral image is normalized to size

32x32 after finding the bounding box of each handwritten numeral image.

## 2.6 Thinning

Thinning is a morphological operation that is used to remove selected foreground pixels from binary images, somewhat like erosion or opening. It can be used for several applications, but is particularly useful for skeletonization

# 3. PROPOSED DESIGN FLOW

The work presented in this paper focuses on recognition of isolated handwritten numerals in Devanagari and Gurumukhi script. In this work the main part is the feature extraction and the classification. We have used four feature extraction techniques like Zoning density, Projection histograms, Distance profiles and Background Directional distribution (BDD). On the basis of these four types of features we have formed 10 feature vectors using different combinations of four basic features which are used for classification. We have used SVM classifier for classification.

## 3.1 Feature Extraction

Feature extraction is extracting information from raw data which is most relevant for classification purpose and that minimizes the variations within a class and maximizes the variations between classes[10]. Selection of a feature extraction method is an important factor in achieving high recognition performance in character recognition systems. The major goal of feature extraction is to extract a set of features, which maximizes the recognition rate with the least amount of confusion. Some of the feature extraction techniques used in our work are discussed below

### *3.1.1 Zonning*

In zoning, the character image is divided into several overlapping or non overlapping zones[6] From each zone features are extracted to form the feature vector. In our work the complete 32×32 numeral image is divided into 16 zones. Hence each zone consist of a total of 64 pixels. Zoning density is obtained by dividing the number of foreground pixels with the total number of pixels in each zone

### *3.1.2 Projection histogram*

Projection histograms count the number of pixels in specified direction. We have used three types of histogram like horizontal, vertical and diagonal. [11] In this the number of foreground pixels are calculated in horizontal, vertical and diagonal direction.

### *3.1.3 Distance Profile*

In distance profile the number of pixels from the bounding box of character are being calculated. we have used profiles of four sides left, right, top and bottom.

### *3.1.4.Background Directional Distribution*

For these features we have considered the directional distribution of neighbouring background pixels to foreground pixels [11] We computed 8 directional distribution features. To calculate directional distribution values of background pixels for each foreground pixel, we have used the masks for each direction

## 3.2 Classification

Classification stage, is the main decision making stage of an OCR system and uses the features extracted in the previous stage to identify the characters . SVM classifier is used for the classification of devanagari and Gurumukhi numerals.

### *3.2.1 SVM classifier*

Support Vector Machine[2] are based on statistical learning theory that uses supervised learning. In supervised learning , a machine is trained instead of programmed, to perform a given task on a number of input-output pairs[11]. 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. The figure 2 shows the classification of objects having class one of the two: either *triangle* or *diamond*.

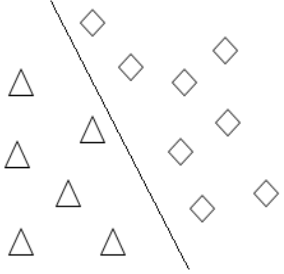

**Figure 4:Linear classification of objects by SVM into two classes**

The separating line defines a boundary on the right side of which all objects are diamond and to the left of which all objects are triangle. Any new object falling to the right is labeled, i.e., classified, as diamond (or classified as triangle if it falls to the left of the separating line). Classification tasks based on drawing separating lines to distinguish between objects of different class memberships are known as hyper plane classifiers. Support Vector Machines are particularly suited to handle such tasks

### *3.2.2 Similarity measure based classifiers*

The distance metric can be termed as similarity measure, which is the key component in content-based image retrieval. Some of such classifiers are discussed below:

#### *3.2.2.1 Euclidean or L2 metric*

It calculates the best which the best distance metric for content based image retrieval. If **x** and y are two d-dimensional feature vectors of database image and query image respectively, then[21] Euclidean or **L2** metric are defined as

$$d\,\mathrm{E}(x,y) = \sqrt{\sum_{i=1}^{d}(x_i - y_i)^2}$$

#### *3.2.2.3 Square chord distance metric*

If **x** and y[21] are two d-dimensional feature vectors of database image and query image respectively then square chord metric are defined as:

$$d\,sc(x,y) = \sum_{i}^{d}(\sqrt{x_i} - \sqrt{y_i})^2$$

# 4. RESULTS

An annotated sample image database of isolated handwritten numerals in Gurumukhi script and Devanagri script has been prepared.In our work a total of 2000 samples of numerals are taken for Gurumukhi and Devanagari. One-fifth of the samples are used for training purposes and four-fifth are used for testing purpose The recognition accuracy obtained by using different combinations of feature vectors for Gurumukhi and Devanagri Numerals are given in the figure below

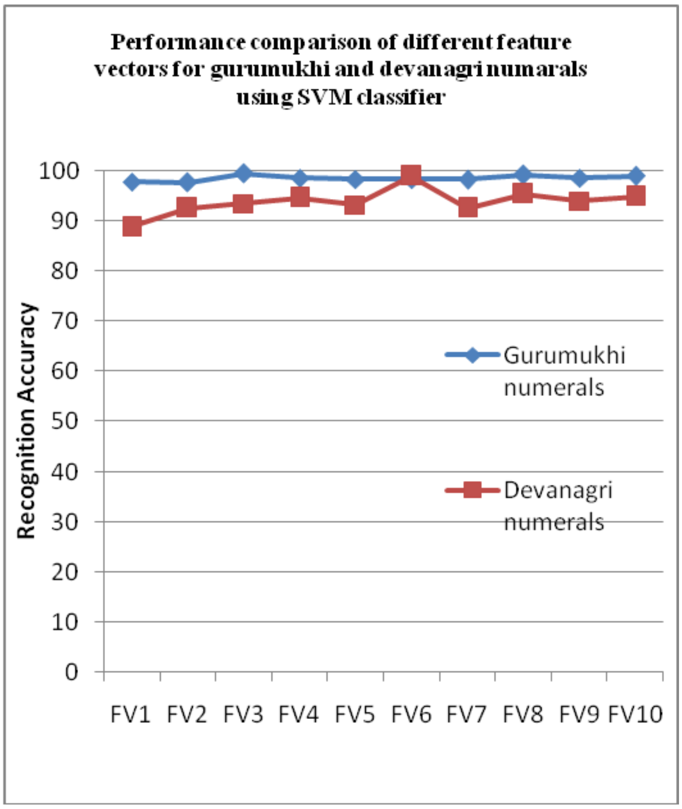

**Figure 5: Comparison of Recognition Accuracy of different different feature vectors for Gurumukhi and Devanagari numerals using SVM classifier**

**Table 1: Performance comparison of Recognition Accuracy different for different feature vectors Gurumukhi and Devanagari numerals using SVM classifier**

| Feature vectors | Gurumukhi Numerals | Devanagri Numerals |
|---|---|---|
| FV1 | 97.80 | 88.80 |
| FV2 | 97.73 | 92.53 |
| FV3 | 99.60 | 93.46 |
| FV4 | 98.60 | 94.60 |
| FV5 | 98.33 | 93.20 |
| FV6 | 98.33 | 99 |
| FV7 | 98.26 | 92.53 |
| FV8 | 99.20 | 95.33 |
| FV9 | 98.60 | 93.93 |
| FV10 | 98.93 | 94.86 |

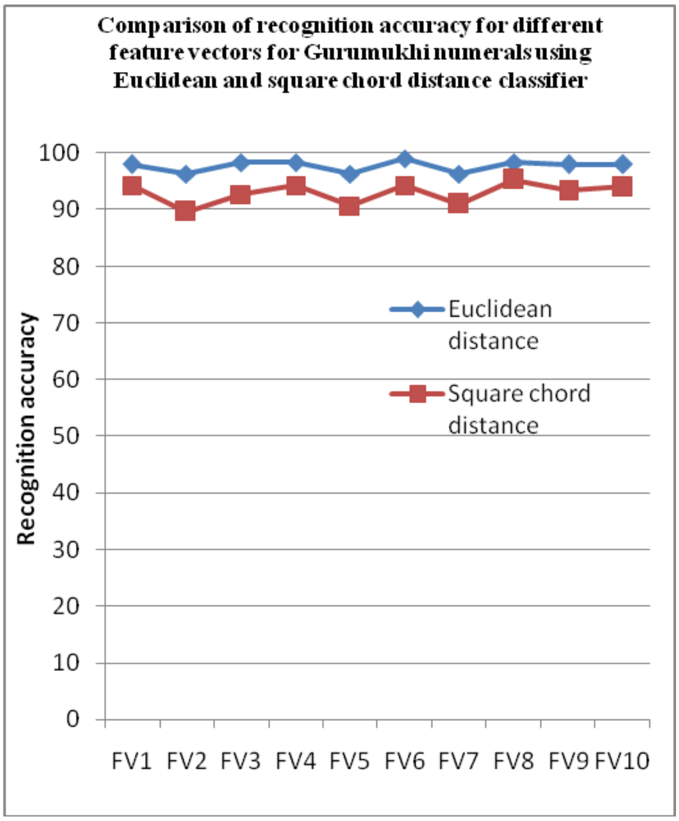

**Figure 6: Comparison of Recognition Accuracy of different different feature vectors for Gurumukhi numerals using Euclidean distance and Square chord distance classifier**

**Table 2: Comparison of Recognition Accuracy of different different feature vectors for Gurumukhi numerals using Euclidean distance and Square chord distance classifier**

| Feature vectors | Using Euclidean distance | Using Square chord distance |
|---|---|---|
| FV1 | 98 | 94.33 |
| FV2 | 96.33 | 89.66 |
| FV3 | 98.33 | 92.66 |
| FV4 | 98.33 | 94.33 |
| FV5 | 96.33 | 90.66 |
| FV6 | 99 | 94.33 |
| FV7 | 96.33 | 91 |
| FV8 | 98.33 | 95.33 |
| FV9 | 98 | 93.33 |
| FV10 | 98 | 94 |

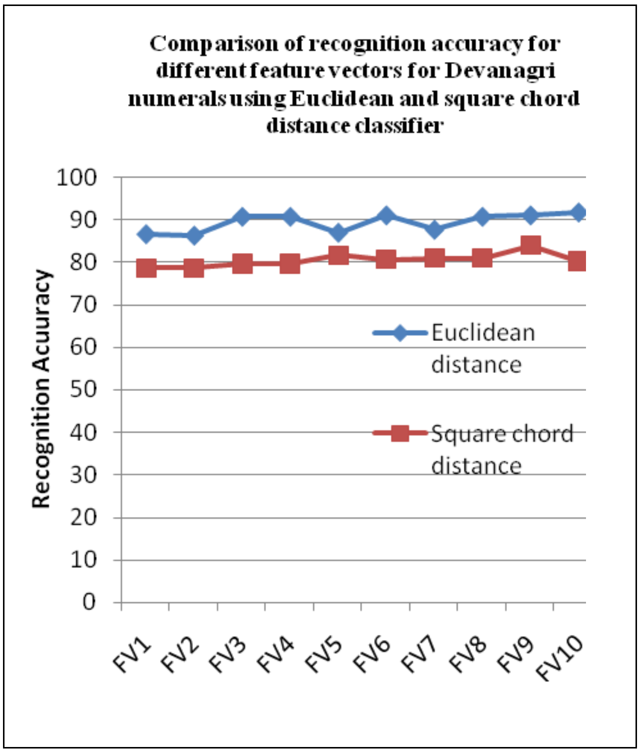

**Figure 7: Comparison of Recognition Accuracy of different different feature vectors for Devanagri numerals using Euclidean distance and Square chord distance classifier**

**Table 3: Comparison of Recognition Accuracy of different different feature vectors for Devanagri numerals using Euclidean distance and Square chord distance classifier**

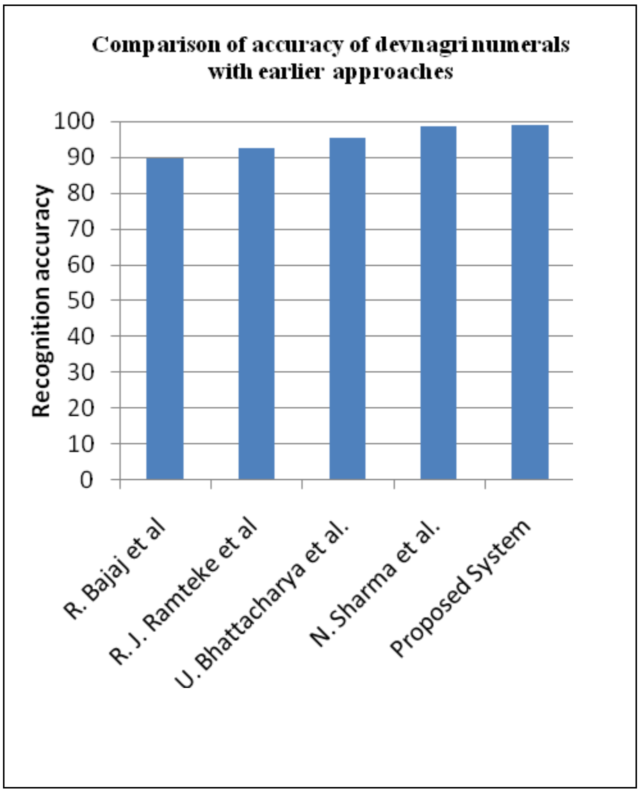

**Table 4: : Comparison of Recognition Accuracy of Devanagri numerals with earlier approaches**

| S.No | proposed by | Accuracy Obtained (%) |
|---|---|---|
| 1 | R. Bajaj et al [12] | 89.6 |
| 2 | R. J. Ramteke et al [13] | 92.68 |
| 3 | U. Bhattacharya et al. [14] | 95.64 |
| 4 | N. Sharma et al. [15] | 98.86 |
| 5 | Proposed System | 99 |

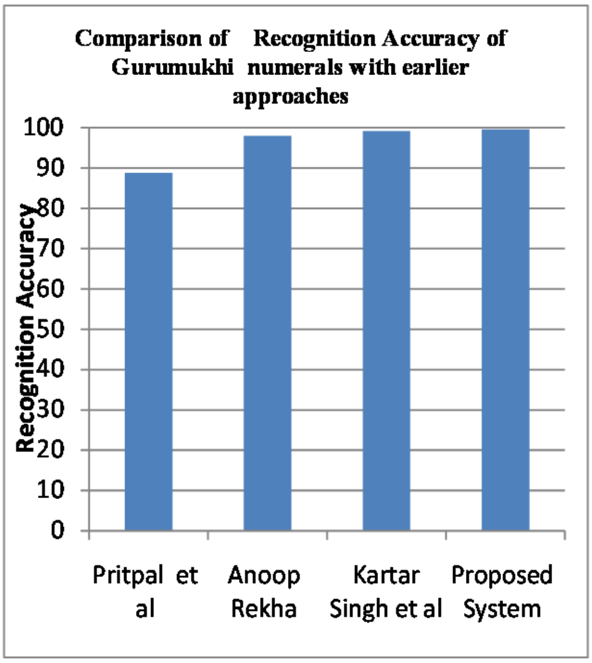

**Figure 9: Comparison of Recognition Accuracy Gurumukhi Numerals with earlier approaches**

**Table 5: Comparison of Recognition Accuracy Gurumukhi Numerals with earlier approaches**

## 5. CONCLUSION

In this paper, we have presented feature extraction and

| S.No | proposed by | Accuracy Obtained (%) |
|---|---|---|
| 1 | Pritpal et al [16] | 88.83 |
| 2 | Anoop Rekha [17] | 98 |
| 3 | Kartar Singh et al. [18] | 99.13 |
| 4 | Proposed System | 99.60 |

classification schemes for optical character recognition of Gurumukhi and devanagari numerals.. For the validation of our result, we compared our result with reported data and significant improvement are observed. We have attain a maximum recognition accuracy of 99.6% in case of Gurumukhi numerals for feature vector FV3 and 99% in devanagri numeral for feature vector FV6 as given in the table 1. The work can be extended to increase the recognition accuracy by adding some more relevant features.

## 6. REFERENCES

[1] Li Lei , Zhang Li-liang, Su Jing-fei,“ Handwritten character recognition via direction string and nearest neighbor matching” The Journal of China Universities of Posts and Telecommunications, pp 160–165 October 2012

[2] Jomy John, Pramod K. V., Kannan Balakrishnan, “ Unconstrained Handwritten Malayalam Character Recognition using Wavelet Transform and Support vector Machine Classifier” International Conference on Communication Technology and System Design pp 598-605 ,2011

[3] Dharamveer Sharma and Puneet Jhajj “Recognition of Isolated Handwritten Characters in Gurumukhi Script” International Journal of Computer Applications Volume 4– No.8, pp 09-17, August 2010

[4] Muhammad Imran Razzak S. A. Hussain Muhammad Sher “ Numeral Recognition for Urdu Script in Unconstrained environment” International Conference on Emerging Technologies pp 44-47, 2009

[5] S. Chanda and U. Pal “English, Devnagari and Urdu Text Identification” Proceedings of the International Conference on Cognition and Recognition*”*

[6] Vikas.j. Dongre and Vijay.H.Mankar “ A review of research on devanagari character recognition” International journal of computer applicationsvolume 12 November 2010

[7] Sandhya Arora, Debotosh Bhattacharjee, Mita Nasipuri, “Combining Multiple Feature Extraction Techniques for Handwritten Devnagari Character Recognition” IEEE Region 10 Colloquium and the Third ICIIS, Kharagpur, INDIA December 8-10, 2008

[8] Malik Waqas Sagheer, Chun Lei He, Nicola Nobile, Ching Y. Suen “Holistic Urdu Handwritten Word Recognition Using Support Vector Machine” 2010 International Conference on Pattern Recognition.

[9] Anil k. .jain and Torfinn Taxt “ Feature Extraction method for character recognition-A Survey ”Elsvier Science Pattern Recognition vol 29 no. 4 pp 641- 662 1996

[10] Anita Rani ,Rajneesh Rani ,Renu Dhir “Combination of Different Feature Sets and SVM Classifier for Handwritten Gurumukhi Numeral Recognition” International Journal of Computer Applications Volume 47– No.18, June 2012

[11] C. J. C. Burges, “A tutorial on support vector machines for pattern recognition”,Knowledge Discovery and Data Mining, Vol. 2(2), pp. 121-167, 1998

[12 Reena Bajaj, Lipika Day, Santanu Chaudhari, “Devanagari Numeral Recognition by Combining Decision of Multiple Connectionist Classifiers”, Sadhana, Vol.27, Part-I, 59-72, 2002

[13] R.J.Ramteke, S.C.Mehrotra, “Recognition Handwritten Devanagari Numerals”, International journal of Computer processing of Oriental languages, 2008

[14] U. Bhattacharya, S. K. Parui, B. Shaw, K. . Bhattacharya, “Neural combination of ANN and HMM for devanagri numeral recognition.”

[15] U. Pal, T. Wakabayashi, N. Sharma and F. Kimura, “Handwritten Numeral Recognition of Six Popular Indian Scripts”, Proc. 9th ICDAR, Curitiba, Brazil, Vol.2 (2007), 749-753

[16] Singh, Pritpal, and Sumit Budhiraja. "Offline handwritten Gurumukhi Numeral Recognition using Wavelet Transforms." *International Journal of Modern Education and Computer Science (IJMECS)* 4.8 (2012)

[17] Rekha, Anoop. "Offline Handwritten Gurmukhi Character and) Numeral Recognition using Different Feature Sets and Classifiers-A Survey."*International Journal of Engineering* (2012.

[18] Siddharth, Kartar Singh, Renu Dhir, and Rajneesh Rani. "Handwritten Gurmukhi Numeral Recognition using Different Feature Sets." *International Journal on Computer Applications*. *(IJCA)* 28.2 (2011): 20-24

[19] Mohamed Cheriet, Nawwaf Kharma, Cheng-Lin Liu, Ching Y. Suen, “Character Recognition Systems: A Guide for Students and Practioners”, *Wiley Inter-Science*, 2007

[20] Manimala Singlia and K.Hemacllandran “Performance analysis of Color Spaces III Image Retrieval” Assam University Journal of Science & Technology: Physical Sciences and Technology Vol. 7 Number II pp-94-104 ,2011

[21] Manesh Kok'are, B.N. Chatterji and **P.K.** Biswas “Comparison of Similarity Metrics for Texture Image Retrieval” IEEE pp 571-575, 2003

[22]Zhang, P., T. D. Bui, and C. Y. Suen. "Hybrid feature extraction and feature selection for improving recognition accuracy of handwritten numerals."*Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on*. IEEE, 2005.