

# Zone Based Feature Extraction and Statistical Classification Technique for Kannada Handwritten Numeral Recognition

Ashoka H.N.

Research Scholar, Department of Electronics & Communication Engineering  
Sikkim Manipal Institute of Technology, Sikkim Manipal University,  
Sikkim, India.

Manjaiah D.H.

Professor & Chairman,  
Department of Computer Science, Mangalore University,  
Mangalore, India.

Rabindranath Bera

Professor, Head of the Department, Department of Electronics & Communication Engineering  
Sikkim Manipal Institute of Technology, Sikkim Manipal University,  
Sikkim, India.

**Abstract**—This paper presents the zone based feature extraction and statistical classification technique for Kannada handwritten numeral recognition. The Kannada handwritten numeral database required for the experimentation is collected from the different individuals and are preprocessed for feature extraction. The binary images fitted in a particular window size are partitioned into a number of regions and a real value is computed by the density of one's to represent the image. The statistical classifier build for the experimentation and on Kannada handwritten numeral database found better classification and recognition rate.

**Keywords**-Handwritten Kannada numeral data; Feature extraction; statistical classifier; Feature matching technique

## I. INTRODUCTION

Kannada is the most popular language of south Indian state Karnataka. It has separate numeral symbols and used extensively to write PIN code, bank cheques, vehicle, and street numbers in the state. Due to their numerous applications, automated reading and recognition of handwritten Kannada numerals finds intensive research. This is the problem of off-line handwritten recognition and is a complex task due to intra, and inters personal writing style variations like size, shape, etc. Researchers proposed different pattern recognition systems based on preprocessing, feature extraction, selection, and classification techniques. The complexity involved in the recognition system is reduced by extracting less number of features to represent the input pattern.

The feature extraction technique is an important step in the pattern recognition system. It extracts a set of relevant information associated to represent the input pattern [1-4]. The quality features such extracted from the samples should have minimum intra-class variation and maximum inter-class variations. The commonly used Feature extraction methods for pattern recognition are statistical, global transformation and series expansion, geometrical and topological techniques [5, 6]. The zoning technique of statistical method is extensively used in the Indian languages handwritten character and numeral recognition for feature extraction [7].

U. Pal et al [7] extracted a feature vector of length 100 for handwritten Kannada numeral recognition by the zoning and directional chain code features and achieved reasonably high accuracy. Rajashekaradhy and Ranjan [8] proposed zone based feature extraction system for four popular south Indian handwritten numeral scripts Kannada, Telugu, Tamil, and Malayalam. They divided binary image of a character into 50 zones and extracted two features from each zone by the zone centroid and image centroid based distance metric. In total, they extracted 100 features to represent the image and reported a recognition rate 95% - 99% by the nearest neighbor classifier (NNC), and feed forward back propagation neural network (BPNN) classifiers for different scripts. In their other work [9], they used support vector machine classifier is used along with NNC and BPNN classifier for Kannada handwritten numeral recognition and obtained a recognition rate of 95%, 92.85% and 96.05% respectively.

Hanumandlu and Murthy [10] extracted 24 features for handwritten numerals representation by the zoning technique and measured coordinate vector distance of 1s pixel from the bottom left corner of the image. They reported the overall recognition rate of 95% for Hindi and 98.4% for English numerals from the fuzzy technique.

Dinesh Acharya et.al [11] extracted five different structural features such as 10-segment string, water reservoir, horizontal/vertical strokes, end points and average boundary length from the minimal bounding box for recognition of isolated handwritten Kannada numerals. They reported a recognition accuracy of 98 % from the extracted 44 potential features by the multiple fuzzy k-NN classifiers. In [12], they reported a recognition rate of 99.5% with 74 features and combined classifiers. Dhanrda et.al [13] extracted potential structural features and used probabilistic neural network classifier for the recognition of Kannada, Telugu and Devanagari numerals. They reported very good recognition rate of 99.40% for Kannada numerals. In other work [14], they divided binary image of a digit into 64 zones and computed pixel density for as feature for bilingual Kannada and Telugu digits recognition. They reported recognition accuracy of 95.5% and 99.83% for Kannada and Telugu numerals respectively with the zoning features and KNN classifier.

From the literature survey, it is clear that many researchers attempted to recognize handwritten numeral digits of Indian scripts. However, they extracted more numbers of features to represent digits and this increases the complexity of classifier. In view of this, an attempt is made to extract less number of features by the zone technique for Kannada handwritten numerals representation and recognition.

## II. METHODOLOGY

The methodology adopted in the work for Kannada handwritten numeral recognition is given in Fig.1. The Kannada handwritten numeral database required for the experimentation is collected from different individuals are binarized and build numeral binary database. Then the database is divided into training and testing data. The training data is used to construct the knowledge bases. Testing data is used to evaluate the performance of the proposed system by the statistical classifiers and fuzzy reasoning classifiers.

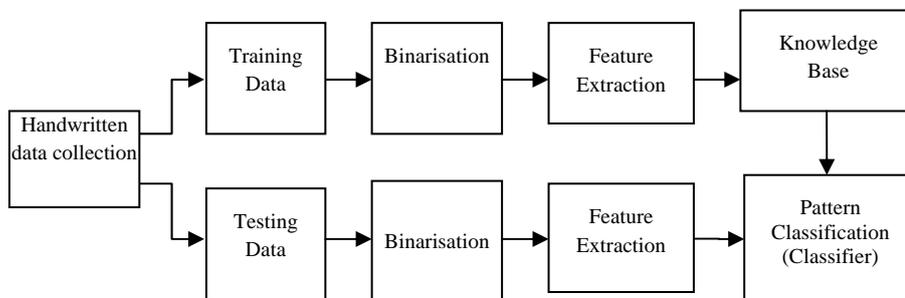


Fig.1 Proposed methodology of Kannada handwritten numeral recognition system

### A. Data collection and Pre-Processing

Since the standard numeral database of south Indian Kannada script is neither available freely or commercially [15]. In view of this, data required for the experimentation is collected from 125 different writers has 1250 training and testing digits and 2500 numerals. Each writer was provided two A4 sheets having ten squares of size 8mm x 8mm in a row one for training and another for testing. Writers were requested to write '0' to '9' Kannada numerals on both the sheets and are scanned using HP-scan jet 5400c at 300 dpi. The blackness introduced in the background of scanned sheet is reduced by increasing the image brightness. Numerals on the scanned sheet are segmented and stored in the separate JPEG format files. Fig.2 shows a set of segmented Kannada handwritten numeral images.

These numeral images are binarized and stored in the window of size 12 x 12 by the built-in functions of MATLAB software. Fig.3 shows the first binary image of Kannada numerals 0-9. The number on top of each binary image is to identify them and in that, the last digit indicates the numeral class and other digits give their serial number in the list.

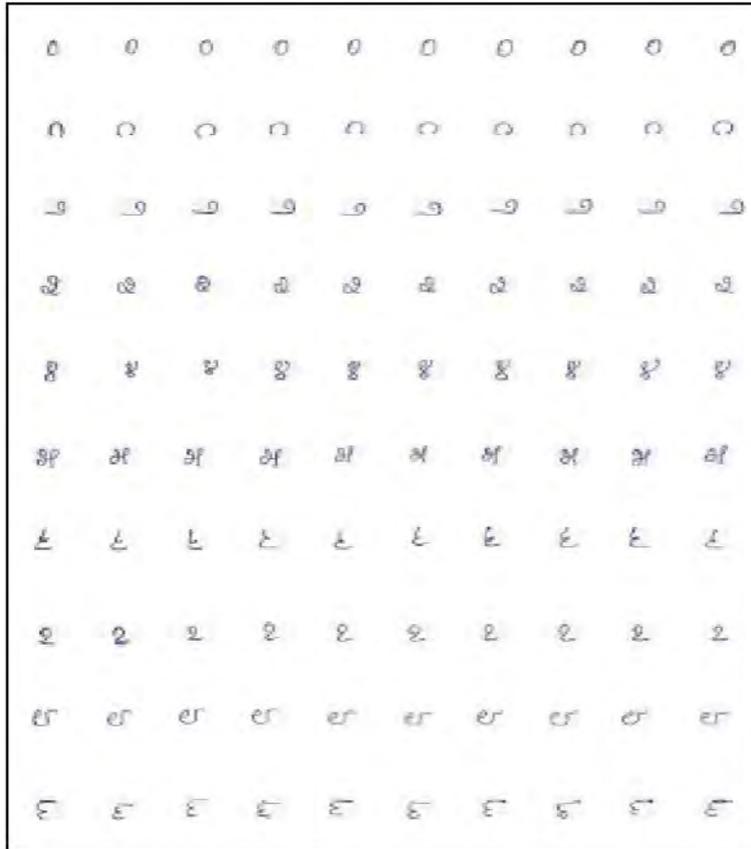


Fig.2 Segmented handwritten Kannada handwritten numeral images.

```

10:      11:      12:      13:      14:
000111111100 001111111110 000000001110 011111111100 001100000001
010000000010 010000000010 00000010001 100100000010 010010000001
100000000010 100000000001 00000010001 111100000001 100010000001
100000000001 100000000001 00000011111 001000000001 011100000010
100000000001 100000000001 00000000001 000000000010 001000000100
100000000001 100000000001 00000000001 000000000010 000100001000
100000000001 100000000001 00000000001 000000000100 000011110000
100000000001 100000000001 00000000001 011111101000 000011100000
100000000010 100000000001 00000000001 010000110000 000110001000
100000000010 100000000001 00000000001 001111101000 000100001000
010000000010 100000000001 000000000010 000000000100 000100010000
001111111100 100000000001 111111111100 000000000010 000011110000

15:      16:      17:      18:      19:
001110000000 00111000000 001111111000 011100000000 001111111111
010001001110 01000000000 010001000100 100100011111 010000000000
010101001001 10000000000 100010000100 100100010000 100000000000
011001001001 10000000000 011100000010 111000001000 010000000000
000001001001 01000000000 000000000010 110000001000 001000000000
000001000110 01111000000 000000000010 100000001000 001111100000
011111111110 01100000000 000000000100 100000001000 001111000000
100001000100 10000000000 000000001000 100000001000 010000000000
100010000100 10000000000 01111110000 100000001000 110000000000
011100000100 00000000000 10000000000 010000001000 100000000000
000000000100 10000000000 10000000000 010000010000 010000000000
000000000100 01111111111 01111111111 001111100000 001111000000
    
```

Fig.3. Binary images of Kannada numerals 0-9.

**B. Feature extraction**

Feature extraction is to represent the input pattern to the pattern classifier. It plays an important role in the pattern recognition system. The zone based feature extraction technique of statistical method used to extract features. In this method, images fitted in a window of particular size are portioned into a number of regions and the black run of a sample or 1s of the binary image is transformed to a real value. In this work, the numeral image fitted in a window of size 12x 12 is partitioned into 16 overlapped sub regions like horizontal, vertical, square, and rectangular for feature extraction as shown in the Fig.4 Most of these regions have the portion of 1s and there could be some regions without 1s and such regions will have zero feature values.

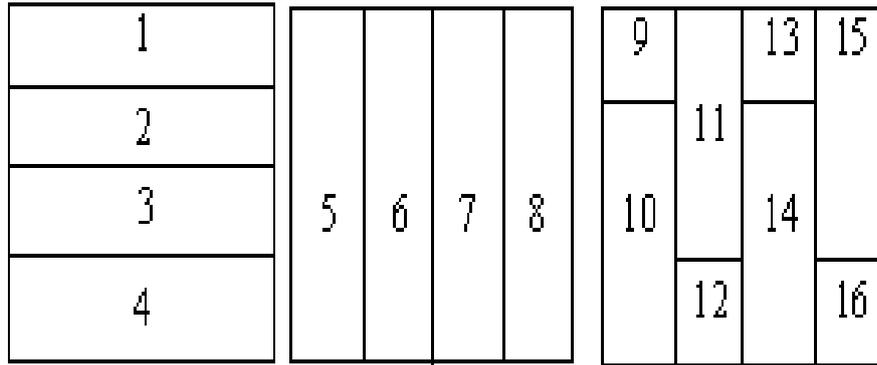


Fig.4: Sub regions Partitioned of binary image.

The feature value of a region computed by the density of one's present in each region. A normalized feature value found for each region by dividing the number of 1s by the total number of elements in the region. The feature value of rth region,  $f_r$  is given by the Eq. (1). This depends on the number of 1s present and not on their position in the region. So the shift in 1s position due to writing styles do not alter the feature value. This increases the robustness of feature extraction method towards the writing style variations.

$$f_r = \frac{1}{q_r} \sum_{i=1}^{q_r} p_{ij} \quad j=1, 2, \dots, n \quad (1)$$

where,  $f_r$  = feature value of rth region,  $q_r$  = total number of elements in the rth region.  $p_{ij}$  = element value in  $i^{th}$  row  $j^{th}$  column of a matrix p .

### III. CLASSIFIER DESIGN AND RECOGNITION

The statistical information, mean feature value and the standard deviation involved in the training sample is used to construct the Knowledge Base (KB). In order to recognize the unknown samples, their feature vectors are matched with the KB and the class with which more number of features matches is assigned to the unknown samples.

#### A. Classifier design

The KB is constructed from the training samples feature vector. The mean feature value ( $\bar{f}_r^k$ ) of each region is computed by the Eq. (3) and the standard deviation ( $\sigma_r^k$ ) by the Eq.(4) for all the numeral classes.

$$\bar{f}_r^k = \frac{1}{m^k} \sum_{i=1}^{m^k} f_{ri}^k \quad (3)$$

$$\sigma_r^k = \sqrt{\frac{1}{m^k} \sum_{i=1}^{m^k} (f_{ri}^k - \bar{f}_r^k)^2} \quad (4)$$

for,  $r = 1, 2, \dots, n$ ; number of regions  
 $i = 1, 2, \dots, m^k$

where,  $m^k$  = number samples in  $k^{th}$  class,  $\bar{f}_r^k$  = mean feature value of rth ,  $f_{ri}^k$  =  $k^{th}$  class  $r^{th}$  region feature value of  $i^{th}$  sample, and  $\sigma_r^k$  = standard deviation of  $k^{th}$  class  $r^{th}$  region feature.

In order to construct the regions KB, minimum and maximum feature value for each region is fixed by adding and subtracting the product of deviator ' $\alpha$ ' and the standard deviation from the region mean feature value. For each region, the feature range for  $k^{th}$  class  $r^{th}$  region is found by the minimum and maximum values and is given by the Eq. (5) and (6) respectively.

$$f_{r \min}^k = \bar{f}_r^k - \alpha \sigma_r^k \quad (5)$$

$$f_{r \max}^k = \overline{f_r^k} + \alpha \sigma_r^k \quad (6)$$

The reference feature regions for each numeral class is constructed and is considered as the KB. To clutch the variability, a deviator ' $\alpha$ ' in the above equation is computed experimentally after evaluating the classification rate.

#### B. Classification and Recognition

To classify the unknown sample, feature vector of it is matched with the feature vectors of all the numeral classes. The unknown sample features are compared with the corresponding reference class region features to measure the similarity. If they are in the reference feature range, then considered features are matched and assigned feature matching index (FMI). If the unknown sample features are in the range of reference class regions a FMI one is given to that region and if not assigned a zero for that region. The general expression for kth class rth region,  $FMI_r^k$  is given in Eq.(7) and total number of features matched with the individual class by the Eq.(8)

$$FMI_r^k = \begin{cases} 1, & \text{if } (f_{r \min}^k \leq f_r^k \leq f_{r \max}^k) \\ 0, & \text{otherwise} \end{cases} \quad (7)$$

$$FMI^k = \sum_{j=1}^n FMI_j^k \quad (8)$$

The class having highest FMI is found and then that is assigned to the unknown sample. In case, if more than one class having highest FMI value, it indicates that the unknown sample matches with all those classes. Even if the unknown sample is recognized as the correctly along with other class is considered as an error.

#### IV. RESULTS AND DISCUSSION

In this section, the experimental results of Kannada handwritten numeral database are presented. For each of the numeral samples, extracted the feature vectors by the method discussed in the section 2.2 and computed the mean feature vectors of each numeral class by the region features. Fig.4 shows the mean feature vectors obtained for the 10 numeral classes. From the Fig.4, it is observed that for numeral classes the mean feature values of each region are different. This indicates that the extracted features are distinctive to represent the individual class. The deviation of each region feature from the mean is found to compute the standard deviation. Fig.5 shows the standard deviation of each region for numeral classes '0-9'.

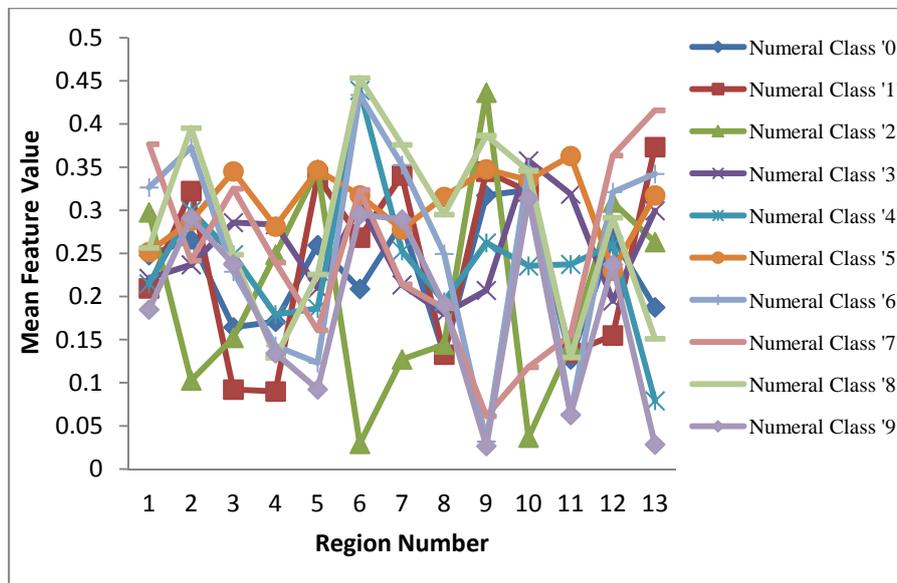


Fig.4. Mean feature vectors of 10 numeral classes

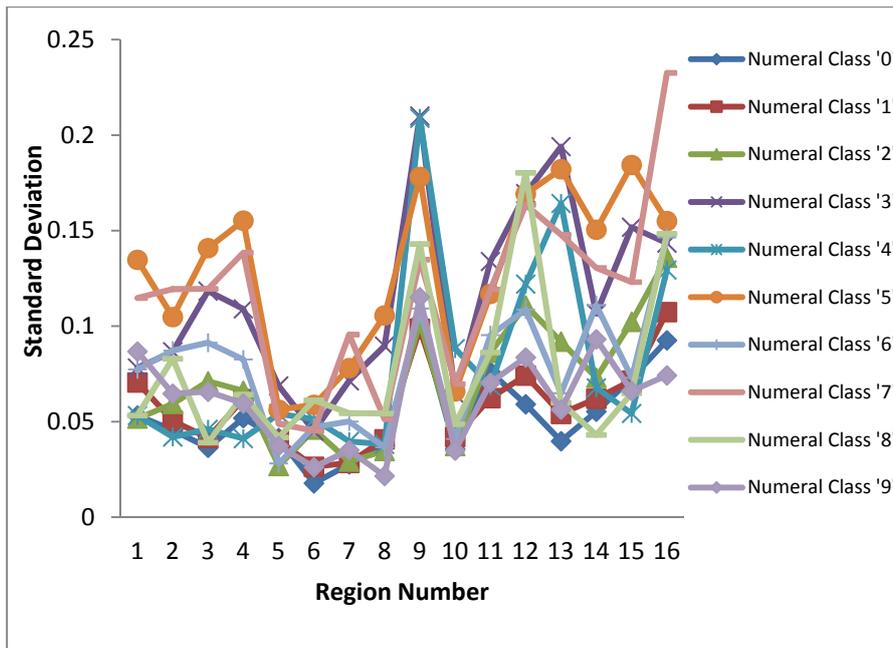


Fig.5. Standard deviation vectors of 10 numeral classes.

The standard deviation is comparatively high for most of the regions. This indicates that, the variability involved among the writers. In order to consider this variability, build the KB for each regions numeral classes using the corresponding standard deviation  $\sigma_r^k$  along with a deviator 'α'. The deviator 'α' found experimentally varying from 0.1 to 3.1 in steps of 0.1. The experimentation is performed on the training samples and found the classification rate. Fig. 6 shows the classification rate obtained for the different deviator 'α'

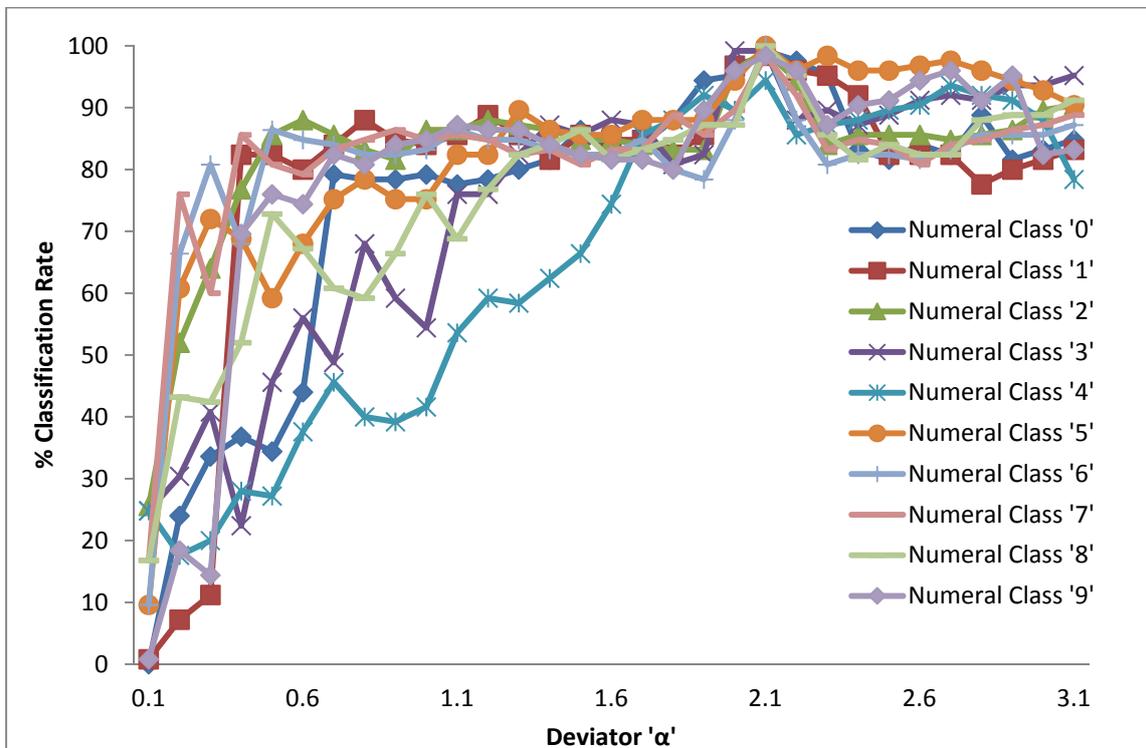


Fig.6: Experimental results with different deviator value α

From the result, it is observed the maximum classification rate at the deviator value  $\alpha= 2.1$ . This value is then used to build the KB and computed the minimum and maximum feature values for each region of all the numeral classes. With this deviator, found the overall classification rate of 99% on training samples. After finding the

feature range, feature vectors of testing samples are extracted similar to the method discussed in section 2.2 and matched with the KB. The number of features lies in the KB is found for each of the numeral class and assigned the FMI. In the experimentation, 125 samples of each numeral class are shown one after the other to the classifier and found the FMI of all the classes. An example, when the 0th numeral class samples are shown to the classifier, the FMI obtained for the first 10 samples are given in the Table 2.

Table.2 FMI of 10 numeral class for the first 10 samples of numeral class '0'.

Sample no.	FMI of numeral class										Unknown sample recognized as
	0	1	2	3	4	5	6	7	8	9	
1	16	13	10	14	14	15	12	14	13	12	0
2	15	13	10	14	13	14	12	14	11	12	0
3	16	14	10	12	13	14	13	12	14	10	0
4	16	12	10	13	12	14	14	13	13	11	0
5	16	12	10	13	12	13	14	13	13	11	0
6	16	13	10	13	13	15	14	13	14	11	0
7	16	13	10	14	14	15	13	14	13	12	0
8	15	11	11	14	12	12	9	14	11	9	0
9	16	12	10	13	12	15	14	13	13	11	0
10	15	13	10	14	13	14	13	13	12	12	0

It is observed that, for most of the samples, the FMI is 16 or 15 for the 0th class and is more among all other numeral classes. However, the FMI of other numeral classes varies from 9 to 15 for some samples. Since the 0th class is having highest FEI value, its class is assigned to all unknown samples. On experimentation with other numeral class unknown samples, same thing is observed and found the overall recognition rate of 98%.

## V. CONCLUSION

From the experimental results, following conclusions are drawn. The feature vector extracted by the zone based technique represent the Kannada handwritten numerals. The variations involved in the interclass mean feature vectors indicate they are distinctive to represent the individual class. The statistical classifier build by the standard deviation is simple and reduced the recognition complexity by extracting less number of features. On experimentation, found better classification rate of 99% with the training samples and recognition rate of 98% with the testing samples.

## References

- [1] Anil K. Jain, Robert P.W. Duin, and Jianchang Mao, "Statistical Pattern Recognition: A Review" IEEE Transactions on Pattern Analysis And Machine Intelligence, Vol. 22, No. 1, 2000, pp. 4-38.
- [2] Husni A. Al-Muhtaseb, Sabri A. Mahmoud and, Rami S. Qahwaji "Recognition of off-line printed Arabic text using Hidden Markov Models", Signal Processing Vol.88,No. 12, 2008 pp. 2902-2912.
- [3] Sabri A. Mahmoud "Recognition of writer-independent off-line handwritten Arabic (Indian) numerals using hidden Markov models." Signal Processing, Vol.88, No. 4, 2008, pp - 844-857.
- [4] G.Y. Chen, W.F. Xie, Pattern recognition with SVM and dual-tree complex wavelets, Image and Vision Computing Vol. 25, 2007, pp 960-966.
- [5] Øivind Due Trier, Anil K. Jain and Torfinn Taxt "Feature Extraction Methods for Character Recognition- A Survey" Pattern Recognition, Vol. 29 No. 4, 1996, pp 641-662.
- [6] Nafiz Arica, Fatos T. Yarman-Vural, "An Overview of Character Recognition Focused on Off-line Handwriting", IEEE Transactions on system Man. Cybernetics-Part C: Applications and Reviews, vol. 31 no. 2 2001, pp. 216-233.
- [7] U. Pal, B. B. Chaudhuri, "Indian Script Character recognition: A survey", Pattern Recognition, vol. 37, 2004, pp 1887-1899.
- [8] Rajashekararadhya S.V. and Vanaja Ranjan P., "Efficient zone based feature extraction algorithm for handwritten numeral recognition of four popular south Indian scripts", Journal of Theoretical and Applied Information Technology, 2005, pp 1171- 1181.
- [9] Rajashekararadhya S.V. and Vanaja Ranjan P., "Handwritten numeral recognition of Kannada script", Proceedings of the international workshop on Machine Intelligence Research (MIR day,GHRCE- Nagpur) 2009, pp-80-86.
- [10] Hanmandlu M., and Murthy O.V.R., "Fuzzy model based recognition of handwritten numerals" Pattern Recognition, vol. 40, 2007 1840 - 1854.
- [11] Dinesh Acharya U, N V Subba Reddy and Krishnamoorthi, "Multilevel classifiers in recognition of Handwritten Kannada numerals", World Academy of Science, Engineering and Technology vol.18, 2008, pp-278-283.
- [12] Dinesh Acharya U, N V Subba Reddy and Krishnamoorthi, "Combined classifiers in recognition of handwritten Kannada numerals: A hybrid approach", International Journal of Information Technology and Knowledge Management, 2009, Vol.2, No. 2 PP 305-311.
- [13] B.V. Dhandra, R.G.Benne and Mallikarjun Hangarge, "Kannada, Telugu and Devanagari handwritten numeral recognition with probabilistic neural network:A novel approach, International Journal of Computer Applications, special issue on RTIPPR-2010, pp 83-88.
- [14] Dhandra B.V., Gururaj Mukarmpi, and Mallikarjun Hangarge, "A script independent approach for handwritten bi-lingual Kannada and Telugu digits recognition" International Journal of Machine Intelligence, Vol.3 No. 3, 2011, PP- 155 -159.
- [15] Benne R.G., Dhandra B.V. and Mallikarjun Hangarge, "Tri-scripts handwritten numeral recognition: a novel approach" Advances in Computational Research, ISSN: 0975-3273, Volume 1, Issue 2, 2009, pp-47-51