HANDWRITTEN NUMERAL/MIXED NUMERALS RECOGNITION OF SOUTH-INDIAN SCRIPTS: THE ZONE-BASED FEATURE EXTRACTION METHOD

[1]S. V. RAJASHEKARARADHYA, [2]Dr P. VANAJA RANJAN

[1]Research Scholar, Department of Electrical and Electronics Engineering, CEG, Anna University

[2]Asst.Professor, Department of Electrical and Electronics Engineering, CEG, Anna University

Chennai, India

Email: [1]svr_aradhya@yahoo.co.in, [2]vanajar@annauniv.edu

ABSTRACT

Handwriting recognition has always been a challenging task in image processing and pattern recognition. There are five major stages in the handwritten character recognition problem: Image processing, segmentation, feature extraction, training and recognition, and preprocessing. India is a multi-lingual multi-script country, where eighteen official scripts are accepted and there are over a hundred regional languages. In this paper we propose a zone-based feature extraction algorithm scheme for the recognition of off-line handwritten numerals of four popular Indian scripts. The character centroid is computed and the character/numeral image (50x50) is further divided into 25 equal zones (10x10). The average distance from the character centroid to the pixels present in the zone column, is computed. This procedure is sequentially repeated for all the zone/grid/box columns present in the zone (10 features). Similarly, the average distance from the character centroid to the pixels present in the zone row is computed. This procedure is sequentially repeated for all the zone rows present in the zone (10 features). There could be some zone column/row that is empty of foreground pixels; then the feature value of that zone column/row in the feature vector is zero. This procedure is sequentially repeated for the entire zone present in the numeral image. Finally, 500 such features are extracted for classification and recognition. The nearest neighbor, feed forward back propagation neural network and support vector machine classifiers are used for subsequent classification and recognition purposes. We obtained a recognition rate of 98.65 % for Kannada numerals, 96.1 % for Tamil numerals, 98.6 % for Telugu numerals and 96.5 % for Malayalam numerals using the support vector machine.

Keywords: *Handwritten Character Recognition, Image processing, Feature Extraction Algorithm, Nearest Neighbor Classifier, Feed forward back propagation neural network, Support Vector Machine, Indian scripts.*

1. INTRODUCTION

Handwriting recognition has always been a challenging task in image processing and pattern recognition. There are five major stages in the handwritten character recognition (HCR) problem: Image preprocessing, segmentation, feature extraction, training and recognition, and post processing. The feature extraction method is probably the most important stage in achieving high recognition performance. HCR has received extensive attention in academic and production fields. The recognition system can be either on-line or off-line. In on-line handwriting recognition, words are generally written on a pressure sensitive surface (digital tablet PCs) from which real time information, such as the order of the stroke made by the writer is obtained and preserved. This is significantly different from off-line handwriting recognition where no dynamic information is available [1]. Off-line handwriting recognition is the process of finding letters and words that are present in the digital image of a handwritten text. It is the subfield of optical character recognition (OCR).

India is a multi-lingual and multi-script country, comprising eighteen official languages, namely, Assamese, Bangla, English, Gujarati, Hindi, Kankanai, Kannada, Kashmiri, Malayalam, Marathi, Nepali, Oriya, Punjabi, Rajasthani,

Vol 7. No 1. (pp 063 - 079)

Sanskrit, Tamil, Telugu and Urdu. Recognition of handwritten Indian scripts is difficult because of the presence of numerals, vowels, consonants, vowel modifiers and compound characters.

Research in HCR is popular for various practical potential applications such as reading aids for the blind, bank cheques, vehicle number plates and automatic pin code reading to sort postal mail. There is a lot of demand on Indian scripts character recognition and a review of the OCR work done on Indian languages is excellently reviewed in [2]. In [3] a survey on the feature extraction methods for character recognition is reviewed. The feature extraction method includes Template matching, Deformable templates, Unitary Image transforms, Graph description, Projection Histograms, Contour profiles, Zoning, Geometric moment invariants, Zernike Moments, Spline curve approximation, Fourier descriptors, the Gradient feature and Gabor features.

We will now briefly review the few important works done towards HCR with reference to the Indian scripts. In [4] for feature computation, the bounding box of a numeral image is segmented into blocks and the directional features are computed in each of the blocks. These blocks are then down-sampled by a Gaussian filter and the features obtained from the down-sampled blocks are fed to a modified quadratic classifier for recognition.

The recognition of isolated handwritten Kannada numerals based on the image fusion method is found in [5]. In [6] the image centroid and zone (ICZ) based distance feature extraction system is proposed. The character centroid is computed and the image is further divided into n equal zones. The average distance from the character centroid to each pixel present in the zone is computed. This procedure is repeated for all the zones present in the numeral image. Finally, n such features are extracted for classification and recognition.

In [7] the zone and vertical downward direction (VDD) projection distance metric feature extraction method is proposed. In [8] the image centroid is computed and the image (50x50) is further divided into 25 equal zones (10x10 each). With reference to the image centroid, the average pixel distance in each zone column in the vertical downward direction is computed (ICZVDD). For each zone 10 features are extracted and this procedure is sequentially repeated for the entire zone present in the image. Finally 250 such features are extracted for classification and recognition.

The selection of the feature extraction method is also an important factor for achieving efficient character recognition. In the previous work of the authors, zone-based features were computed with reference to the image and zone centroid. In the present work, the zone-based feature extraction method is further explored. For the benefit of the readers, the presentation, structure and information are explained briefly in the next sections. Readers can refer to [15] for more detailed information.

In this paper we propose a zone-based feature extraction algorithm. The Nearest neighbor classifier (NNC), Feed forward Back propagation neural network (BPNN) and support vector machine (SVM) classifiers are used for the recognition and classification of a numeral image. We have tested our method for Kannada, Tamil, Telugu and Malayalam numerals individually and also for the mixed numerals.

The rest of the paper is organized as follows. In Section 2 we shall briefly explain the overview of Indian scripts. In Section 3 we shall explain the dataset and preprocessing. In Section 4 we shall discuss the proposed extraction method. Section 5 describes the classifiers, experimental results and comparative study, and finally the conclusion is given in Section 6.

2. BRIEF OVERVIEW OF KANNADA, TAMIL, TELUGU AND MALAYALAM SCRIPTS

Kannada is one of the major Dravidian languages of Southern India, one of the earliest languages evidenced epigraphically in India, and spoken by about 50 million people in the Indian states of Karnataka, Tamil Nadu, Andhra Pradesh and Maharashtra. The script has 49 characters in its alphasyllabary and is phonetic. The script also includes 10 different Kannada numerals of the decimal number system.

Tamil is a Dravidian language, and one of the oldest in the world. It is the official language of the Indian state of Tamil Nadu; it also has an official status in Sri Lanka, Malaysia and Singapore. The Tamil script has 10 numerals, 12 vowels, 18 consonants and five grantha letters.

Telugu is a Dravidian language and has the third most popular script in India. It is the official language of the Indian state of Andhra Pradesh. There are 10 numerals, 18 vowels, 36 consonants, and three dual symbols.

Malayalam is a Dravidian language, and has the eighth most popular script in India, and is spoken by about 30 million people in the Indian state of Kerala. Both the language and the writing system are closely related to Tamil. However, Malayalam has its own script. The script has 16 vowels, 37 consonants and 10 numerals.

The most challenging part of the Indian handwritten character recognition is the distinction between similar shaped components. A very small variation between two characters or numerals leads to recognition complexity and a certain degree of recognition accuracy. The style of writing the characters is highly different, as they come in various sizes and shapes. The same numeral may take different shapes, and conversely, two or more different numerals of a script may take a similar shape.

3. DATASET AND PREPROCESSING

Earlier 2000 Kannada numeral samples from 200 different writers were collected. Writers were provided with a plain A4 sheet, and each writer was asked to write Kannada numerals from 0 to 9 at one time [15]. Recently, we have again collected 2000 Kannada numerals by 40 different writers. In this paper the dataset size of 4000 Kannada numerals is used. The database is totally unconstrained and has been created for validating the recognition system. Similarly, the dataset size of 2000 Tamil numerals, 2000 Telugu numerals and 2000 Malayalam numerals is used. Please note that the previous dataset [15] is also included in the present dataset.

The collected documents are scanned using the HP-scan jet 5400c at 300dpi, which is usually a low noise and good quality image. The digitized images are stored as binary images in the BMP format. A sample of Kannada, Tamil, Telugu, and Malayalam handwritten numerals from the dataset are shown from figures 1 to 4 respectively.

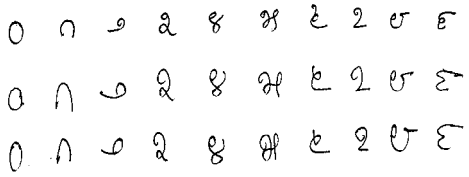

Figure 1. Sample of handwritten Kannada numerals from 0 to 9.

Figure 2 Sample of handwritten Tamil numerals from 0 to 9

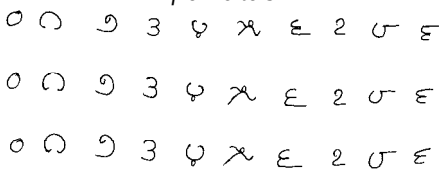

Figure 3. Sample of handwritten Telugu numerals from 0 to 9

Figure 4 Sample of handwritten Malayalam numerals from 0 to 9

The preprocessing stage involves noise reduction, slant correction, size normalization and thinning. Among these, size normalization and thinning are very important. 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 50x50 after finding the bounding box of each handwritten numeral image.

Thinning provides a tremendous reduction in data size; it extracts the shape information of the characters. It can be considered as the conversion of off-line handwriting to almost on-line data. Thinning is the process of reducing the thickness of each line of pattern to just a single pixel.

In this research work, we have used the morphology-based thinning algorithm for better symbol representation. Detailed information about the thinning algorithm is available in [9]. Thus, the reduced pattern is known as the skeleton and is close to the medial axis, which preserves the topology of the image. Figure 5 shows the steps involved in our method as far as preprocessing is considered. Figure 6 shows the preprocessed numeral samples (shown in appendix).

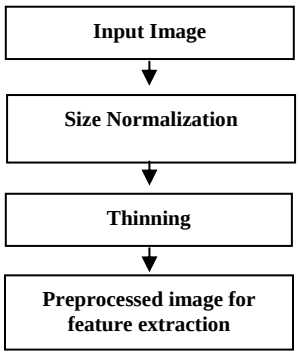

Figure 5. Preprocessing of the input numeral image

4. PROPOSED FEATURE EXTRACTION METHOD

For extracting the feature, the zone-based approach is followed. The zone-based feature extraction method gives good results even when certain preprocessing steps like filtering, smoothing and slant removing are not considered. In this section, we shall explain the concept of the proposed feature extraction method used for extracting features for efficient classification and recognition.

The image size is selected after considerable experimentation on different image sizes and this detail is available in [10]. The character centroid is computed and the character/numeral image (50x50) is further divided into 25 equal zones (10x10) as shown in Figure 7. The average distance from the character centroid to the pixels present in the zone column is computed. This procedure is sequentially repeated for all the zone/grid/box columns present in the zone (10 features). Similarly, the average distance from the character centroid to the pixels present in the zone row is computed. This procedure is sequentially repeated for all the zone/grid/box rows present in the zone (10 features). There could be some zone column/row that is empty of foreground pixels, and then the feature value of that zone column/row in the feature vector is zero. This procedure is sequentially repeated for the entire zone present in the image (500 features).

The algorithm, provided below, shows the general working procedure of our proposed method.

Proposed algorithm

Zone based feature extraction system (ICZVDD-ICZHRD)

ICZVDD refers to the zone column distance feature in the vertical downward direction with reference to the image centroid.
ICZHRD refers to the zone row distance feature in the horizontal right direction with reference to the image centroid.

Input: Preprocessed numeral image

Output: Features for Classification and Recognition

Method Begins

Step 1: Compute the input image centroid
Step 2: Divide the input image into **25** equal zones.
Step 3: Compute the distance between the Image centroid to the pixel present in the zone column.
Step 4: Repeat step 3 for the entire pixel present in the zone column.
Step 5: Compute the average pixel distance in the zone column. (One feature)
Step 6: Repeat steps 3 to 5 for the entire zone column present in the zone (Ten features).
Step 7: Compute the distance between the image centroid to the pixel present in the zone row.
Step 8: Repeat step 7 for the entire pixel present in the zone row.
Step 9: Compute the average pixel distance in the zone row. (One feature)
Step 10: Repeat steps 7 to 9 for the entire zone row present in the zone (Ten features).
Step 11: Repeat the steps 3-10 sequentially for the entire zone present in the image.
Step 12: Finally, 500 such features will be obtained for classification and recognition

Method Ends

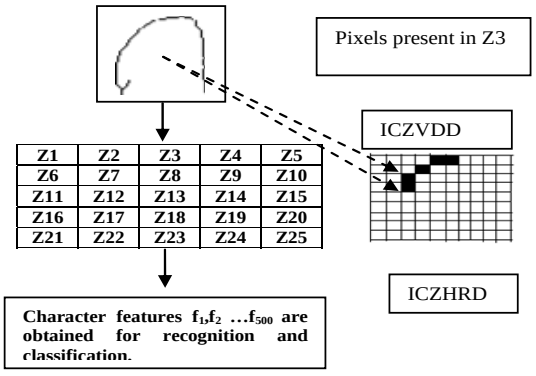

Figure 7. Procedure for extracting features from the numeral image

5. CLASSIFIERS, EXPERIMENTAL RESULTS AND COMPARATIVE STUDY

5.1. Nearest neighbor classifier

For large-scale pattern matching, a long-employed approach is the nearest neighbor classifier. The training phase of the algorithm consists only of storing the feature vectors and class labels of the training samples. In the actual classification phase, the same features as before are computed for the test samples. Distances from the new vector to all the stored vectors are computed. Then Classification and recognition is achieved on the basis of similarity measurement.

5.2. Feed-forward back propagation neural network classifier

An Artificial Neural Network (ANN) is a computational model widely used in pattern recognition. It has been used extensively both for the recognition of non-Indian as well as Indian digits. Recognition of handwritten numerals is a very complex problem. The feed forward back propagation neural network is used for subsequent recognition and classification of numeral images.

The recognition performance of the BPNN will depend largely on the structure of the network and the training algorithm. The number of nodes in the input, hidden and output layers will determine the network structure. All the neurons of one layer are fully interconnected with all the neurons of its just preceding and just succeeding layers (if any). The network consists of 500 nodes in the input layer (corresponding to 500 features). The output layer has 10 neurons corresponding to 10 numerals. Therefore, only the number of nodes in the hidden layer needs to be determined.

The architecture of the BPNN consists of one input layer, one hidden layer and one output layer. The number of hidden nodes will heavily influence the network performance. Insufficient hidden nodes will cause under-fitting, where the network cannot recognize the numeral because there are not sufficient adjustable parameters to model the input-output relationship. Excessive hidden nodes will cause over-fitting where the network fails to generalize. There is no theoretical development, based on which, the optimal number of neurons in the hidden layer can be determined.

There are several rules of the thumb for deciding the number of neurons in the hidden layer [11].

- The number of hidden neurons should be less than twice the input layer size.
- The number of hidden neurons should be between the size of the input layer and the size of the output layer.
- Performance of the training and testing samples.
- The number of hidden neurons should be 2/3 of the input layer size, plus the size of the output layer.

While selecting the number of neurons for the hidden layer, the suitable thumb rule mentioned can be considered individually. The selection of hidden neurons for optimizing the network structure is very important. In the case of the Kannada numeral dataset (total dataset size is 4000 samples and for each class 400 samples are taken) 2000 samples are used for training and 2000 samples for testing purposes.

To select the hidden layer nodes once again 50 % of the training samples are used for learning and the remaining 50 % of the training samples are used for validation. The hidden layer nodes are varied to achieve optimal training and testing sample performance. The performance of the network is shown in Table 1 for the Kannada numeral dataset. We followed these procedures to arrive at 80 neurons for the hidden layer. Again, using all the training samples network structure is retrained and then 2000 testing samples are fed to the network (not included in the training/validation/learning phase) to find the classification and recognition.

Table 1 Network Performance for Different Hidden Layer nodes for Kannada Handwritten Numerals. RR Refers to the Recognition Rate

Hidden layer nodes	Epochs	RR (in %) Training Samples	RR (in %) Testing Samples
30	907	98.3	92.1
40	1742	98.2	92
50	488	97.8	93.4
60	65	97.5	94.1
70	126	97	94
80	85	97.5	**94.3**
90	184	97.4	93.7
100	171	98.2	93.9
110	619	94.9	91.9

Similarly in the case of Tamil, Telugu and Malayalam the numeral dataset (for each script, total dataset size is 2000 samples) are considered. The performance of the network is shown in Tables 2, 3 and 4 for Tamil, Telugu and Malayalam numeral dataset respectively. Hence 60, 40 and 40

hidden layer neurons are selected for Tamil, Telugu and Malayalam numerals respectively.

Table 2 Network Performance for Different Hidden Layer nodes for Tamil Handwritten Numerals. RR Refers to the Recognition Rate

Hidden layer nodes	Epochs	RR (in %) Training Samples	RR (in %) Testing Samples
40	663	99.6	87
50	771	97.2	86
60	657	98.2	**88.6**
70	263	97.4	88
80	973	97	86.8
90	442	95.8	87.8

Table 3 Network Performance for Different Hidden Layer nodes for Telugu Handwritten Numerals. RR Refers to the Recognition Rate

Hidden layer nodes	Epochs	RR (in %) Training Samples	RR (in %) Testing Samples
30	1443	99.4	96.6
40	150	99.8	**97.6**
50	95	99.6	97
60	205	97	94.8
70	173	97.2	95.4
80	1808	93.2	92

Table 4 Network Performance for Different Hidden Layer nodes for Malayalam Handwritten Numerals. RR Refers to the Recognition Rate

Hidden layer nodes	Epochs	RR (in %) Training Samples	RR (in %) Testing Samples
40	368	100	**93.2**
50	256	98	90.6
60	317	91.4	99.8
70	294	96.2	84.6
80	72	99.2	91.4
90	755	96.6	90.6

Since our desired outputs must range between 0 and 1, we have selected the log sigmoid as the transfer function for both the hidden and output layers. We have used the 'Mean Squared Error' (MSE) as the performance parameter function. MSE is the average squared error between the network outputs and the target outputs. During training, the weights of the network are iteratively adjusted to minimize the function. We have adopted the 'Gradient descent back propagation' as a learning algorithm. The algorithm updates weights according to the gradient descent momentum and adaptive learning rate. The training parameters are the learning rate (0.1), momentum factor (0.8) and performance goal (0.01), selected after intensive experimentation.

5.3. Support vector machine classifier

The support vector machine is a new classifier that is extensively used in many pattern recognition applications. The SVM uses the principle of structural risk minimization by minimizing the Vapnik Chervonenkis (VC) dimensions [12, 13]. Regarding the pattern classification problem, the SVM demonstrates very good generalization performance in empirical applications.

The SVM is a binary classifier that separates linearly any two classes by finding a hyper plane of a maximum margin between the two classes. The margin means the minimal distance from the separating hyper plane to the closest data points. The SVM learning machine searches for an optimal separating hyper plane, where the margin is maximal. The outcome of the SVM is based only on the data points that are at the margin and are called support vectors.

There are two approaches to extend the SVM for multi-class classification. The first is the one-against-one approach (ONO), and the other is the one-against-all (ONA). We have used the ONA approach for the decomposition of the learning problem from N class pattern recognition into several two class learning problems. The SVM is constructed for each class by discriminating that class against the remaining (N-1) classes. The recognition based on this approach consists of N SVM models. The set of training samples belongs to the corresponding class is used to build the SVM for a corresponding class. In the case of handwritten digit recognition, ten SVM models are constructed corresponding to the ten classes.

All the training examples (labeled with class) are used to construct the SVM model for a class. The SVM model for class 0 (numeral 0) is constructed using a set of training samples and their desired outputs. The desired output is +1 if the training sample belongs to class 0; otherwise, desired output is -1. Similarly, SVM models are constructed for the remaining classes. The testing sample is classified based on the evidence obtained from each of the SVM models.

The kernel is utilized to map the input data to a higher dimensional feature space, so that the

problem becomes linearly separable. The kernel plays a very important role. The Gaussian kernel performs better compared to the linear kernel, polynomial kernel etc. The Gaussian kernel has been used here.

5.4. Experimental results

In order to evaluate the performance of the proposed method, we shall consider handwritten Kannada, Tamil, Telugu and Malayalam Numerals. For Kannada numerals, 4000 samples are considered. Here, 2000 samples are used for training purposes and the remaining 2000 samples are used for testing. Similarly, 2000 samples are considered for each Tamil, Telugu and Malayalam numerals. Here 1000 numeral samples are considered for training and 1000 numeral samples for testing purposes. For recognition and classification purposes NNC, BPNN and SVM classifiers are used.

Table 5, 6, 7 and 8 give the result for Kannada, Tamil, Telugu and Malayalam handwritten numerals respectively. Also we have used a two fold cross validation (CV) scheme for recognition result evaluation (for NNC and SVM classifiers). Here the dataset of handwritten numerals is divided into 2 subsets and testing is done on each subset, using the rest of the subset for learning. The recognition rates of all the 2 test subsets of the dataset are averaged to get the recognition result.

Table 5 Result of Kannada Handwritten Numerals. CRR Refers to the Class Recognition Rate in %. ORR Refers to the Overall Recognition Rate in %.

For Kannada numerals Training samples = 2000 and testing samples = 2000							
	NNC classifier			BPNN classifier	SVM classifier		
Kannada Numerals	CRR I fold	CRR II fold	CRR (CV)	CRR	CRR I fold	CRR II fold	CRR (CV)
0	100	99.5	99.75	99.5	100	99.5	99.75
1	100	100	100	89.5	100	99	99.5
2	100	100	100	98	100	100	100
3	93	96.5	94.75	94.5	97	99	98
4	98.5	97.5	98	99	99.5	99.5	99.5
5	94.5	99.5	97	94	98	100	99
6	96	97	96.5	95.5	96.5	95.5	96
7	98	92	95	93.5	97	95.5	96.25
8	99.5	96.5	98	98.5	98.5	96	97.25
9	100	99.5	99.75	99	100	99	99.5
ORR	**97.95**	**97.8**	**97.875**	**96.1**	**98.65**	**98.3**	**98.475**

Table 6 Result of Tamil Handwritten Numerals. CRR Refers to the Class Recognition Rate in %. ORR Refers to the Overall Recognition Rate in %.

For Tamil numerals Training samples =1000 and testing samples = 1000							
	NNC classifier			BPNN classifier	SVM classifier		
Tamil Numerals	CRR I fold	CRR II fold	CRR (CV)	CRR	CRR I fold	CRR II fold	CRR (CV)
0	99	99	99	96	98	99	98.5
1	94	92	93	91	97	98	97.5
2	97	96	96.5	92	95	99	97
3	94	99	96.5	98	95	100	97.5
4	100	96	98	95	97	96	96.5
5	99	93	96	98	98	95	96.5
6	92	82	87	90	93	82	87.5
7	99	100	99.5	93	97	99	98
8	97	91	94	97	99	93	96
9	85	87	86	89	91	96	93.5
ORR	**95.6**	**93.5**	**94.55**	**93.9**	**96**	**95.7**	**95.85**

Table 7 Result of Telugu Handwritten Numerals. CRR Refers to the Class Recognition Rate in %. ORR Refers to the Overall Recognition Rate in %.

For Telugu numerals Training samples =1000 and testing samples = 1000							
	NNC classifier			BPNN classifier	SVM classifier		
Tamil Numerals	CRR I fold	CRR II fold	CRR (CV)	CRR	CRR I fold	CRR II fold	CRR (CV)
0	97	98	97.5	97	99	99	99
1	96	99	97.5	97	99	97	98
2	97	97	97	94	98	97	97.5
3	99	95	97	96	97	97	97
4	98	98	98	97	100	98	99
5	89	98	93.5	93	99	99	99
6	100	97	98.5	100	100	98	99
7	97	98	97.5	93	94	99	96.5
8	100	100	100	99	100	100	100
9	100	97	98.5	100	100	97	98.5
ORR	**97.3**	**97.7**	**97.5**	**96.6**	**98.6**	**98.1**	**98.35**

Table 8 Result of Malayalam Handwritten Numerals. CRR Refers to the Class Recognition Rate in %. ORR Refers to the Overall Recognition Rate in %.

For Malayalam numerals Training samples =1000 and testing samples = 1000							
	NNC classifier			BPNN classifier	SVM classifier		
Tamil Numerals	CRR I fold	CRR II fold	CRR (CV)	CRR	CRR I fold	CRR II fold	CRR (CV)
0	98	100	99	95	95	100	97.5
1	99	100	99.5	97	100	100	100
2	99	98	98.5	91	95	99	97
3	93	91	92	94	95	98	96.5
4	91	90	90.5	82	91	96	93.5
5	96	98	97	98	98	100	99
6	100	98	99	98	97	100	98.5
7	97	100	98.5	95	98	96	97
8	100	92	96	100	100	91	95.5
9	97	89	93	92	96	98	97
ORR	**97**	**95.6**	**96.3**	**94.2**	**96.5**	**97.8**	**97.15**

5.5. Experimental results on multi-lingual document (handwritten numerals of south-Indian scripts)

In this paper, the authors also address the issue of the recognition of mixed handwritten numerals of four Indian scripts, namely, Kannada, Tamil, Telugu and Malayalam. In order to evaluate the robustness of the proposed feature extraction algorithm, experiments on a multi-lingual document (south-Indian numerals written on a document belong to different scripts) are also carried out. Handwritten numerals in mixed scripts are frequently found in the Indian postal mail and tabular form documents.

We considered a 1000 (100 samples per class) handwritten numeral samples from each handwritten numeral database of Kannada, Tamil, Telugu and Malayalam scripts as training samples. Hence 4000 samples are used as a training set. Similarly, we considered a 1000 (100 samples per class) handwritten numeral samples from each handwritten numeral database of Kannada, Tamil, Telugu and Malayalam scripts as a testing samples. Hence 4000 samples are used as testing set. Please note that the training and testing set are different. The SVM classifier is used for subsequent classification and recognition. Tables 9, 10, 11 and 12 gives the individual results for Kannada, Tamil, Telugu and Malayalam handwritten numerals respectively with respect to the multi-lingual

(classes too are mentioned) south-Indian numeral document. Also we have used a two fold cross validation (CV) scheme for recognition result evaluation. Here the dataset of multi-lingual handwritten numerals is divided into two subsets and testing is done on each subset using the rest of the subset for learning. The recognition rates of all the two test subsets of the dataset are averaged to get recognition result.

Table 9 Result of Kannada Handwritten Numerals of Multi-Lingual South-Indian Numeral Document. CRR Refers to the Class Recognition Rate in %. ORR Refers to the Overall Recognition Rate.

Class	Kannada numeral	I fold CRR	II fold CRR	CV CRR
Class-0	0	45	43	44
Class-1	1	96	90	93
Class-2	2	81	84	82.5
Class-3	3	90	97	93.5
Class-4	4	97	94	95.5
Class-5	5	90	96	93
Class-6	6	64	76	70
Class-7	7	84	69	76.5
Class-8	8	82	90	86
Class-9	9	68	74	71
ORR in %		**79.7**	**81.3**	**80.5**

Table 10 Result of Tamil Handwritten Numerals of Multi-Lingual South-Indian Numeral Document. CRR Refers to the Class Recognition Rate in %. ORR Refers to the Overall Recognition Rate.

Class	Tamil numeral	I fold CRR	II fold CRR	CV CRR
Class-10	0	40	48	44
Class-11	1	93	97	95
Class-12	2	80	85	82.5
Class-13	3	91	99	95
Class-14	4	98	95	96.5
Class-15	5	87	91	89
Class-16	6	96	84	90
Class-17	7	95	99	97
Class-18	8	98	91	94.5
Class-19	9	93	96	94.5
ORR in %		**87.1**	**88.5**	**87.8**

Table 11 Result of Telugu Handwritten Numerals for Multi-Lingual South-Indian Numeral Document. CRR Refers to the Class Recognition Rate in %. ORR Refers to the Overall Recognition Rate.

Class	Telugu numeral	I fold CRR	II fold CRR	CV CRR
Class-20	0	37	20	28.5
Class-21	1	90	82	86
Class-22	2	63	81	72
Class-23	3	98	96	97
Class-24	4	99	90	94.5
Class-25	5	83	92	87.5
Class-26	6	83	68	75.5
Class-27	7	75	97	86
Class-28	8	88	91	89.5
Class-29	9	67	63	65
ORR in %		**78.3**	**78**	**78.15**

Table 12 Result of Malayalam Handwritten Numerals of Multi-Lingual South-Indian Numeral Document. CRR Refers to the Class Recognition Rate in %. ORR Refers to the Overall Recognition Rate.

Class	Malayalam numeral	I fold CRR	II fold CRR	CV CRR
Class-30	0	83	92	87.5
Class-31	1	99	98	98.5
Class-32	2	88	91	89.5
Class-33	3	91	97	94
Class-34	4	89	93	91
Class-35	5	98	81	89.5
Class-36	6	98	99	98.5
Class-37	7	85	89	87
Class-38	8	100	88	94
Class-39	9	96	98	97
ORR in %		**92.7**	**92.6**	**92.65**

5.6. Comparative results for south-Indian handwritten numerals

Table 13 gives the comparative result for Kannada Numerals related to the I fold (First 50 % samples of each class for training and remaining 50 % samples of each class for testing). Table 14 gives the comparative result for Kannada numerals based on cross validation. The results are tabulated for a common dataset under the same experimental procedure.

Method 1: Here the zone-based feature extraction system is proposed. With reference to the image centroid, the average foreground pixel distance for each zone (ICZ) has been computed [6].

Method 2: Here the zone-based feature extraction system is proposed. With reference to the image centroid and zone centroid the average foreground pixel distance for each zone has been computed. (ICZ-ZCZ) [15].

Table 13 provides a comprehensive comparison for Kannada handwritten numerals using three different classifiers. The recognition result and run time for Kannada handwritten numerals are tabulated for the three classifiers individually (I fold). Similarly, Table 14 provides the recognition

result and run time for Kannada handwritten numerals for the NNC and SVM classifiers individually (cross validation). Also, for the proposed method, we have extracted the feature by dividing the image (50x50) into 4 zones. Hence, for each zone 50 features are extracted. The results provided in Table 14 show that the proposed algorithm provides high recognition rate compared to Method 1 and 2.

Table 15 provides a comprehensive comparison of Tamil handwritten numerals using three different classifiers. The recognition result and run time for Tamil handwritten numerals are tabulated for the three classifiers individually (I fold). Similarly, Table 16 provides the recognition result and run time for Tamil handwritten numerals for the NNC and SVM classifiers individually (cross validation).

Table 17 provides a comprehensive comparison of Telugu handwritten numerals using three different classifiers. The recognition result and run time for Telugu handwritten numerals are tabulated for the three classifiers individually (I fold). Similarly, Table 18 provides the recognition result and run time for Telugu handwritten numerals for the NNC and SVM classifiers individually (cross validation).

Table 19 provides a comprehensive comparison of Malayalam handwritten numerals using three different classifiers. The recognition result and run time for Malayalam handwritten numerals are tabulated for the three classifiers individually (I fold). Similarly, Table 20 provides the recognition result and run time for Malayalam handwritten numerals for the NNC and SVM classifiers individually (cross validation).

Table 13 Comparative results for Kannada Handwritten Numerals (I fold). ORR Refers to the Overall Recognition Rate.

Kannada handwritten numerals (I fold) Training samples=2000 & Testing samples= 2000						
	NNC classifier		BPNN classifier		SVM classifier	
Method	RR (In %)	Run time (Seconds)	RR (In %)	Run time (Seconds)	RR (In %)	Run time (Seconds)
Method 1-ICZ	96.95	53.0113	93.8	70.6094	96.95	4.57
Method 2- ICZ-ZCZ	97.3	71.8438	94.2	88.2969	97.1	98.51
Proposed method ICZVDD-ICZHRD	**97.95**	230.4844	**96.1**	151.6563	**98.65**	**63.25**

Table 14 Comparative results for Kannada Handwritten Numerals (Cross Validation). ORR Refers to the Overall Recognition Rate.

Kannada handwritten numerals - Cross validation				
	NNC classifier		SVM classifier	
Method	RR (In %)	Run time (Seconds)	RR (In %)	Run time (Seconds)
Method 1-ICZ	96.625	55.1932	96.60	4.535
Method 2-ICZ-ZCZ	97.125	72.3356	96.975	17.755
Proposed method ICZVDD-ICZHRD	**97.875**	230.6407	**98.475**	56.37
Proposed method ICZVDD-ICZHRD 4-zones	**97.15**	**109.9375**	**97.55**	**33.575**

Table 15 Comparative results for Tamil Handwritten Numerals (I fold). ORR Refers to the Overall Recognition Rate

Tamil handwritten numerals (I fold) Training samples=1000 & Testing samples= 1000						
	NNC classifier		BPNN classifier		SVM classifier	
Method	RR (In %)	Run time (Seconds)	RR (In %)	Run time (Seconds)	RR (In %)	Run time (Seconds)
Method 1-ICZ	90.9	13.4531	90.2	87.1250	91.3	2
Method 2- ICZ-ZCZ	92.4	18.0469	91.2	101.7188	92.3	5.85
Proposed method ICZVDD-ICZHRD	**95.6**	53.75	**93.9**	57.1875	**96.1**	23.63

Table 16 Comparative results for Tamil Handwritten Numerals (Cross Validation). ORR Refers to the Overall Recognition Rate.

Tamil handwritten numerals - Cross validation				
	NNC classifier		SVM classifier	
Method	RR (In %)	Run time (Seconds)	RR (In %)	Run time (Seconds)
Method 1-ICZ	91.15	13.4375	91.3	2.1
Method 2-ICZ-ZCZ	92	17.4922	92	5.795
Proposed method ICZVDD-ICZHRD	**94.55**	54.9186	**95.9**	35.045

Table 17 Comparative results for Telugu Handwritten Numerals (I fold). ORR Refers to the Overall Recognition Rate

Telugu handwritten numerals (I fold) Training samples=1000 & Testing samples= 1000						
	NNC classifier		BPNN classifier		SVM classifier	
Method	RR (In %)	Run time (Seconds)	RR (In %)	Run time (Seconds)	RR (In %)	Run time (Seconds)
Method 1-ICZ	93.3	12.5156	92.3	22.9219	94.5	2.84
Method 2- ICZ-ZCZ	93.9	12.0469	92.6	28.6713	94.7	5.52
Proposed method ICZVDD-ICZHRD	**97.3**	54.8906	**96.6**	61.2031	**98.6**	17.04

Table 18 Comparative results for Telugu Handwritten Numerals (Cross Validation). ORR Refers to the Overall Recognition Rate.

Telugu handwritten numerals - Cross validation				
	NNC classifier		SVM classifier	
Method	RR (In %)	Run time (Seconds)	RR (In %)	Run time (Seconds)
Method 1-ICZ	94.35	12.9062	95.05	3.87
Method 2-ICZ-ZCZ	94.9	17.4687	95.3	5.5
Proposed method ICZVDD-ICZHRD	**97.5**	54.5078	**98.35**	16.555

Table 19 Comparative results for Malayalam Handwritten Numerals (I fold). ORR Refers to the Overall Recognition Rate.

Malayalam handwritten numerals (I fold) Training samples=1000 & Testing samples= 1000						
	NNC classifier		BPNN classifier		SVM classifier	
Method	RR (In %)	Run time (Seconds)	RR (In %)	Run time (Seconds)	RR (In %)	Run time (Seconds)
Method 1-ICZ	95.2	11.5781	93.4	30.9063	96.2	3.93
Method 2- ICZ-ZCZ	95.8	16.4688	93.4	55.3438	96.6	5.54
Proposed method ICZVDD-ICZHRD	**97**	54.2031	**94.2**	31.8281	**96.5**	28.52

Table 20 Comparative results for Malayalam Handwritten Numerals (Cross Validation). ORR Refers to the Overall Recognition Rate.

Malayalam handwritten numerals - Cross validation				
	NNC classifier		SVM classifier	
Method	RR (In %)	Run time (Seconds)	RR (In %)	Run time (Seconds)
Method 1-ICZ	94.75	11.6328	95.15	2.925
Method 2-ICZ-ZCZ	95.3	16.4141	95.55	4.485
Proposed method ICZVDD-ICZHRD	**96.3**	55.1406	**97.15**	22.945

5.7. Comparative results for south-Indian mixed handwritten numerals

Table 21 gives the comparative result for mixed numerals. The first 100 samples of each class of all the south-Indian numerals are considered from training set (4000 samples). Similarly, 100 samples of each class of all the south-Indian numerals are considered from the testing set (4000 samples). Please refer to section 5.6 for Method 1 and Method 2.

The Cross validation results are tabulated individually for Kannada, Tamil, Telugu and Malayalam handwritten numerals (Mixed handwritten numeral document). The results are averaged to get the recognition result for mixed numerals. We obtained 84.775 % for mixed south-Indian handwritten numerals.

Table 21 Comparative results for south-Indian mixed handwritten Numerals (Cross Validation). ORR Refers to the Overall Recognition Rate.

South-Indian mixed handwritten numerals - Cross validation using SVM classifiers					
	Kannada numerals	Tamil numerals	Telugu numerals	Malayalam numerals	Mixed Numerals
Method	ORR (In %)	ORR (In %)	ORR (In %)	ORR (In %)	ORR (In %)
Method 1-ICZ	72.3	79.4	71.2	87	77.48
Method 3-ICZ-ZCZ	72.55	80.3	73.05	87.8	78.425
Proposed method ICZVDD-ICZHRD	**80.5**	**87.8**	**78.15**	**92.65**	**84.775**

5.8. Discussions

Handwritten numeral recognition (HNR) plays an important role in various applications. It is a subset of the more general category of optical character recognition and it has evolved to be a class of its own, due to its myriad application potential. India is a multi-lingual multi-script country, where eighteen official scripts are accepted and there are over a hundred regional languages. Handwriting recognition in the Indian context is an active research topic and very few works are found on Indian scripts. In this direction we have contributed an efficient feature extraction method in this paper.

We have conducted and tabulated the results related to south-Indian handwritten numerals (in section 5.6). Here each handwritten numeral script is individually trained and tested (single script document).

We also extended the experimental analysis for mixed numerals. In section 5.7 the results are tabulated for south-Indian mixed handwritten numerals (multi-lingual document).

5.9. Experimental result on the MNIST database

In this section, we experimentally evaluate the performance of the proposed method on well-known MNIST database of handwritten digits (http://yann.lecun.com/exdb/mnist).

The MNIST database consists of 60,000 training samples and 10,000 testing samples. All the digits have been size normalized and centered in a 28–by–28 box. Preprocessed samples of the MNIST handwritten digits (here size normalization (50x50) and thinning are performed) are shown in Fig 8 (shown in appendix). We have considered 5000 training samples and 1000 testing samples for our experimental analysis. For the proposed feature extraction method, we achieved a **97.2 %** recognition rate using the SVM classifier.

5.10. Experimental result on the ISI Bangla numeral database

In this section, we experimentally evaluate the performance of the proposed method on the well-known ISI Bangla numerals database [14]. The ISI Bangla numeral database has 19,392 training samples and 4000 test samples. Preprocessed samples of the ISI Bangla handwritten digits (here size normalization (50x50) and thinning are performed) are shown in Fig 9 (shown in appendix). We have considered 12000 training samples and 3000 testing samples for our experimental analysis. For the proposed feature extraction method, we achieved a **95.47 %** recognition rate using the SVM classifier. Table 22 provides the summary of a few previous methods and result for handwritten Bangla numerals using the proposed method.

Table 22 Summary of some existing results and the result achieved with the proposed method.

Ref	Script	Feature	Classifier	#Train	#Valid	#Test	Recognition rate
[4]	Bangla	Direction	Quadratic	14,650	Cross validation	Cross validation	98.99
[16]	Bangla	Wavelet	MLP, MV	6000	1000	5000	97.16
[17]	Bangla	Kirsh, PCA	SVM, MV	6000	0	10000	95.05
Proposed method	Bangla	Zone	SVM	12000	0	3000	95.47

Pal et al. have reported the results of handwritten numeral recognition of six popular Indian scripts, including Bangla [4]. For feature computation, the bounding box of a numeral image is segmented into blocks and the directional features are computed in each of the blocks. These blocks are then down-sampled by a Gaussian filter and the features obtained from the down-sampled blocks are fed to a modified quadratic classifier for recognition. By a five fold cross validation (rotation) on 14,650 Bangla numeral samples, they reported a recognition rate of 98.99 %.

Bhattacharya and Chaudhuri have proposed a multi-resolution wavelet analysis and a majority voting approach for the handwritten numeral recognition of the Bangla script [16]. They have extracted the wavelet filtering features in three resolutions and the multi-layer perceptron (MLP) neural network for the classification on each resolution. Then, the classification results were combined by a majority vote (MV). They have used 6000 Bangla numeral samples for training, 1000 samples for validation and 5000 samples for testing. They achieved a recognition rate of 97.16 %.

Wen et al. have proposed two approaches for handwritten Bangla numeral recognition [17]. The first one is based on image reconstruction error in principal subspace and the second one is based on the Kirsh gradient (four masks for four orientations); dimensionality reduction is achieved by principal component analysis (PCA) and classification is achieved by the SVM. A third approach is based on image PCA and SVM classification. The results of all these three recognizers are combined by a majority vote. They have used 6000 Bangla numeral samples for training and 10,000 samples for testing. They have achieved a recognition rate of 95.05 %.

6. CONCLUSION

In this paper we have proposed a zone-based projection distance feature extraction algorithm for the recognition of four popular Indian numeral scripts. The nearest neighbor, feed forward back-propagation neural network and support vector machine classifiers are used for subsequent classification and recognition. We have obtained a maximum recognition rate of 98.65 % for Kannada handwritten numerals using the support vector machine classifier. We also extended the experimental analysis for south-Indian mixed numerals and the results are encouraging.

In this work we have considered 6000 images from the MNIST database and 15000 images from the ISI Bangla numeral database.

Extensive experiments on different numeral databases of Indian scripts and standard database like the MNIST and the ISI Bangla, show that the proposed zone-based feature extraction system is robust.

One can observe that the sample of preprocessed images contains pruning images, disconnected images, slanted images etc. The preprocessing stage used is very simple and involves only size normalization and thinning, and classification is achieved with well known classifiers like NNC, BPNN, and SVM. Our main contribution involves feature computation. We intended to extend our work further to contribute new zone-based feature extraction techniques for efficient classification and recognition purposes to the research community.

In addition, we plan to extend our work to other Indian scripts. Also, the effective use of the hybrid classifier concept to achieve a good recognition rate is one of our future research directions.

REFERENCES

[1] R. Plamondon and S. N. Srihari, "On-line and off- line handwritten character recognition: A comprehensive survey", *IEEE. Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 63-84, 2000.

[2] U. Pal, B. B. Chaudhuri, "Indian Script Character recognition: A survey", *Pattern Recognition*, vol. 37, pp. 1887-1899, 2004.

[3] Anil.K.Jain and Torfinn Taxt, "Feature extraction methods for character recognition-A

Survey", *Pattern Recognition*, vol. 29, no. 4, pp. 641-662., 1996,

[4] U. Pal, T. Wakabayashi and F. Kimura, "Handwritten numeral recognition of six popular scripts", *Ninth International conference on Document Analysis and Recognition ICDAR 07*, Vol.2, pp.749-753, 2007.

[5] G.G. Rajaput and Mallikarjun Hangarge, "Recognition of isolated handwritten Kannada numerals based on image fusion method: ", *PReMI07*, LNCS.4815, pp.153-160, 2007.

[6]S.V. Rajashekararadhya and P. Vanaja Ranjan, "Isolated handwritten Kannada digit recognition: A novel approach", *Proceedings of the International Conference on Cognition and Recognition"*, pp.134-140, 2008.

[7] S.V. Rajashekararadhya, P. Vanaja Ranjan and V.N. Manjunath Aradhya, "Isolated handwritten Kannada and Tamil numeral recognition: A novel approach", *First International Conference on Emerging Trends in Engineering and Technology ICETET 08*, pp.1192-1195, 2008.

[8] S.V. Rajashekararadhya, and P. Vanaja Ranjan ,"Handwritten numeral recognition of Kannada and Telugu scripts: ", *International Conference on emerging trends in computing*, pp.165-169, 2009.

[9] Rafael C. Gonzalez, Richard E. woods and Steven L. Eddins, *Digital Image Processing using MATLAB*, Pearson Education, Dorling Kindersley, South Asia, 2004.

[10] S.V. Rajashekararadhya, and P. Vanaja Ranjan ,"Support vector machine based handwritten numeral recognition of Kannada script:", *IEEE International advance computing Conference*, pp.381-386, 2009

[11] A. Majumdar and B.B. Chaudhuri, "Printed and handwritten Bangla numeral recognition using multiple classifier outputs", *Proceedings of the first IEEE ICSIP06*, Vol. 1, pp. 190-195, 2006

[12] V. N. Vapnik, *Statistical Learning Theory*, John Wiley and sons, 1998.

[13] V. N. Vapnik, *The nature Statistical Learning Theory*, Springer, New York, 2nd edition, 1999.

[14] U. Bhattacharya, and B. B. Chaudhuri, " Handwritten numeral databases of Indian scripts and multistage recognition of mixed numerals", *IEEE Transaction on Pattern analysis and machine intelligence*, vol.31, No.3, pp.444-457, 2009.

[15] S.V. Rajashekararadhya, and P. Vanaja Ranjan ,"Efficient zone based feature extraction algorithm for handwritten numeral recognition of four popular south-Indian scripts:", *Journal of Theoretical and Applied Information Technology*, *JATIT* vol.4, no.12, pp.1171-1181, 2008.

[16] U. Bhattacharya, and B. B. Chaudhuri, " A majority voting scheme for multiresolution recognition of hand-printed numerals:", *In proceedings of the 7th International conference on document analysis and recognition ICDAR 2003*, Edinburgh, Scotland, pp.789-793, 2005.

[17] Y. Wen, Y. Lu, and P. Shi, " Handwritten Bangla numeral recognition system and its application to postal automation: ", *Pattern Recognition*, vol.40, no.1, pp.99-107, 2007.

APPENDIX

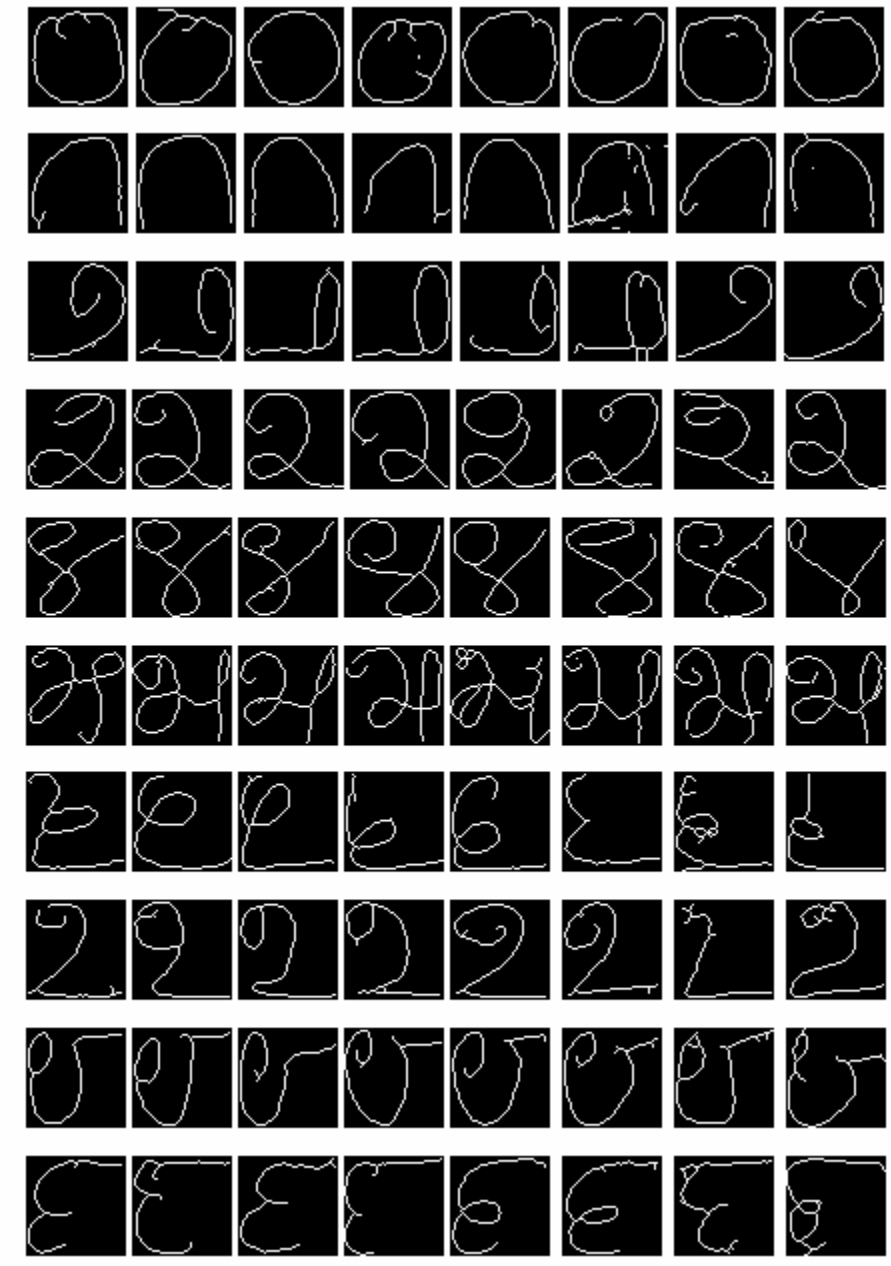

Figure 6. Sample of preprocessed images of handwritten numerals from the Kannada digit database

APPENDIX

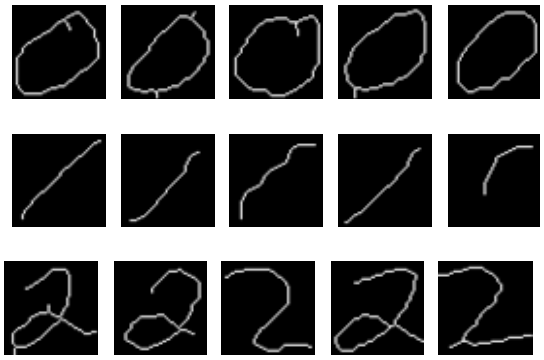

Figure 8. Sample of preprocessed images of handwritten numerals from the MNIST digit database

Vol 7. No 1. (pp 063 - 079)

APPENDIX

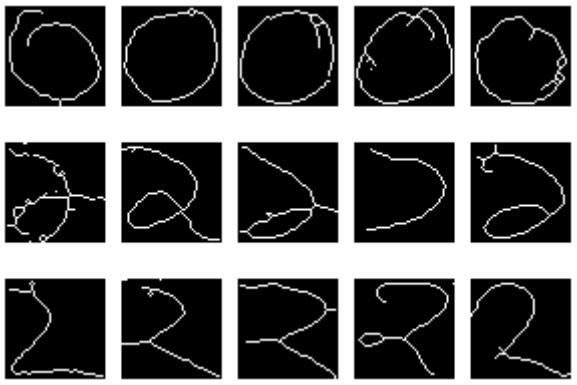

Figure 9. Sample of preprocessed images of handwritten numerals from the ISI Bangla digit database