Handwritten Numeral Recognition of Kannada Script

S.V. Rajashekararadhya
Department of Electrical and Electronics Engineering
CEG, Anna University, Chennai, India
svr_aradhya@yahoo.co.in

P. Vanaja Ranjan
Department of Electrical and Electronics Engineering
CEG, Anna University, Chennai, India
vanajar@annauniv.edu

Abstract

Handwritten character recognition has received extensive attention in academic and production fields. The recognition system can be either online or off-line. There is a large demand for handwritten character recognition and hand written documents. India is a multi-lingual and multi script country, where eighteen official scripts are accepted and have over hundred regional languages. In this paper we present the zone based angle feature extraction system. The numeral image centroid is computed and the image is further divided into n equal zones. Average angle from the character centroid to the pixels present in the zone is computed. This procedure is repeated sequentially for all zones present in the numeral image. Finally n such features are extracted. For classification and recognition purpose, nearest neighbor classifier and support vector machines are used. We achieved 96.05 % of recognition accuracy using support vector machines for Kannada numerals.

1. Introduction

Handwritten character recognition (HCR) has received extensive attention in academic and production fields. HCR is the important area in image processing and pattern recognition. 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 to off-line handwriting recognition where no dynamic information is available [1]. Off-line handwriting recognition is the process of finding letters and words are present in digital image of handwritten text. It is the subfield of optical character recognition (OCR).

There are five major stages in the HCR problem: Image preprocessing, segmentation, feature extraction, training and recognition and post processing. Research in HCR is popular for various practical application potential such as reading aid for the blind, bank cheques, vehicle number plate, automatic pin code reading of postal mail to sort. There is a lot of demand on Indian scripts character recognition and a review of the OCR work done on Indian language is excellently reviewed in [2]. In [3] a survey on feature extraction methods for character recognition is reviewed. 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 , Gradient feature and Gabor feature.

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

We will now briefly review the few important works done towards HCR with reference to the Kannada and Telugu scripts. In [4] for feature computation, the bounding box of a numeral image is segmented into blocks and the directional features have 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 have fed to a modified quadratic classifier for recognition. Recognition of isolated handwritten Kannada numerals based on image fusion method is available in [5]. The numeral has normalized to fit into a size of 32x32 pixels. The binary numeral image is divided into 64 zones of equal size, each zone being of size 4x4 pixels.

Then the total foreground pixels of the image have computed. Sequentially total pixels of each zone have computed. If any zone total pixel density greater than 5% of total foreground pixels, 1 is stored for that particular zone, other wise 0 is stored.

In [6] Zone and Distance metric based feature extraction is used. The character centroid has computed and the image is further divided in to n equal zones. Average distance from the character centroid to the each pixel present in the zone has computed. This procedure has repeated for all the zones present in the numeral image. Finally n such features are extracted for classification and recognition. In [7] zone and vertical projection distance metric has used. The preprocessed numeral image (50x50) is fed to the feature extraction module. The image is divided into 25 zones (each zone size is 10x10). Average pixel distance of the each column present in the zone has computed vertically. Hence 10 distance features are obtained for zone. This procedure is repeated sequentially for all the zones. Finally 250 features are extracted recognition.

In [8] zone centroid is computed and the image is further divided in to n equal zones. Average distance from the character centroid to the each pixel present in the zone has computed. This procedure is repeated for all the zones present in the numeral image. Finally n such features are extracted for classification and recognition. Neural network based handwritten Kannada numeral recognition is found in [9]. Selection of feature extraction method is also a most important factor for achieving efficient character recognition.

The rest of the paper is organized into five sections. In Section 2 we will briefly explain about the overview of Kannada script. In Section 3 we will briefly explain about the data collection and preprocessing. In Section 4 we will discuss about the feature extraction method. Section 5 describes about classifies, experimental results and comparative study and finally, conclusion is given in Section 6.

2. Overview of Kannada Script

In this section, we will explain the properties of Kannada script, which is one of the popular South Indian scripts. Most of the Indian scripts are originated from Brahmi script through various transformations. Writing style of Indian scripts considered in this paper is from left to right, and the concept of upper/lower case is not applicable to these scripts.

Kannada is one of the major Dravidian languages of Southern India and one of the earliest languages evidenced epigraphically in India and spoken by about 50 million people in the Indian state of Karnataka, Tamil Nadu, Andra Pradesh and Maharasthra. The script has 49 characters in its alphasyllabary and is phonetic. The characters are classified into three categories: swaras(vowels), vyanjans(consonants) and yogavaahas (part vowel, part consonants). The scripts also include 10 different Kannada numerals of the decimal number system.

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

3. Data collection and preprocessing

Data collection for the experiment has been done from the different individuals. Currently we are developing dataset for Kannada and Telugu numeral scripts. We have collected 4000 Kannada numeral samples from 140 different writers. Writers were provided with the plain A4 sheet and each writer has asked to write Kannada numerals from 0 to 9 for one time. The database is totally unconstrained and has been created for validating the recognition system. The collected documents are scanned using 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 BMP format.

A sample of Kannada handwritten numerals from data set are shown in Figure. 1.

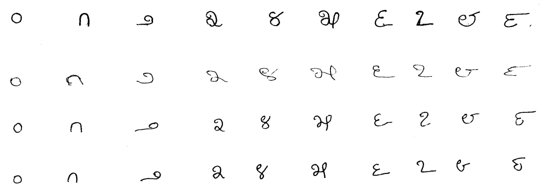

Figure 1. Sample handwritten Kannada numerals 0 to 9

Preprocessing includes the steps that are necessary to bring the input data into an acceptable form for feature extraction. The raw data, depending on the data acquisition type, is subjected to a number of preliminary processing stages. 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, thinning extracts the shape information of the characters. It can be considered as conversion of off-line handwriting to almost on-line data. Thinning is the process of reducing thickness of each line of pattern to just a single pixel. In this research work, we have used morphology based thinning algorithm for better symbol representation. The detail information about the thinning algorithm is available in [11]. Thus the reduced pattern is known as the skeleton and is close to the medial axes, which preserves the topology of the image. Figure. 2 shows the steps involved in our method as par preprocessing is considered.

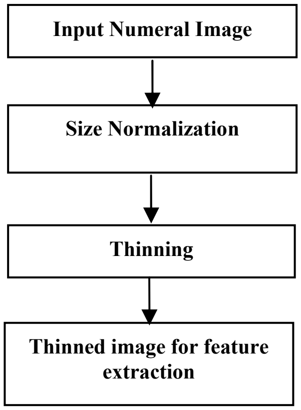

Figure 2. Preprocessing of the input numeral image

4. Feature extraction method

The most important aspect of handwriting recognition scheme is the selection of good feature set, which is reasonably invariant with respect to shape variations caused by various writing styles. Zone based feature extraction method provides good result even when certain preprocessing steps like filtering, smoothing and slant removing are not considered. For extracting the feature, approach presented in [10] is used here. In this section, we explain the concept of feature extraction algorithm used for extracting the features for efficient classification and recognition. Nearest neighbor classifier (NNC), Feed forward back propagation neural network (BPNN) and support vector machine (SVM) are used as classifiers. The following paragraph explains in detail about the feature extraction methodology.

The numeral image centroid is computed and the character/numeral image (50x50) is further divided in to 50 equal zones as shown in Fig. 3. Average angle from the character centroid to pixels present in the zone is computed. This procedure is repeated sequentially for the entire zone present in the numeral image. There could be some zone having empty foreground pixels, then the value of that particular zone in the feature vector is zero. Finally 50 such features are extracted for classification and recognition.

The graphical representation about the logic is shown in Figure. 3. For classification and recognition nearest neighbor classifier is used. The following is the algorithm to show the working procedure of the proposed feature extraction method.

Algorithm: Image centroid and Zone based Angle (ICZA) feature extraction system.

Input: Preprocessed handwritten numeral image

Output: Features for Classification and Recognition

Method Begins

Step 1: Compute the input image centroid

Step 2: Divide the input image in to **n** equal zones.

Step 3: Compute angle between the image centroid to pixel present in the zone.

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

Step 5: Compute average angle between these points. (one feature)

Step 6: Repeat the steps 3 to 5 sequentially for the entire zone present in the image.

Step 7: Finally, **n** such features are obtained for classification and recognition.

Method Ends

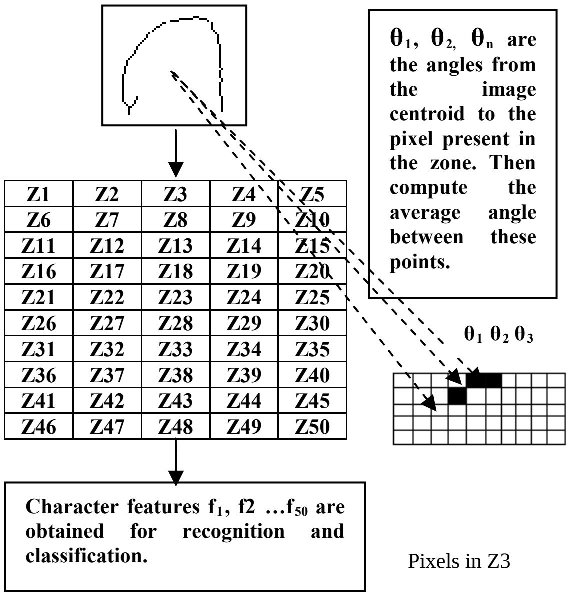
θ₁, θ₂, θₙ are the angles from the image centroid to the pixel present in the zone. Then compute the average angle between these points.
Z1 Z2 Z3 Z4 Z5
Z6 Z7 Z8 Z9 Z10
Z11 Z12 Z13 Z14 Z15
Z16 Z17 Z18 Z19 Z20
Z21 Z22 Z23 Z24 Z25
Z26 Z27 Z28 Z29 Z30
Z31 Z32 Z33 Z34 Z35
Z36 Z37 Z38 Z39 Z40
Z41 Z42 Z43 Z44 Z45
Z46 Z47 Z48 Z49 Z50
θ₁ θ₂ θ₃
Character features f₁, f2 …f₅₀ are obtained for recognition and classification.
Pixels in Z3

Figure 3. Procedure for extracting features from the numeral image

5. Classifiers, experimental results and comparative study

5.1. Nearest neighbor classifier for classification and recognition

For large-scale pattern matching, a long-employed approach is the NNC. The training phase of the algorithm consists only of storing the feature vectors 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 stored vectors are computed. Then Classification and recognition is achieved on the basis of similarity measurement.

5.2. Feed forward back propagation neural network for classification and recognition

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 numeral is a very complex problem. BPNN is used for subsequent recognition and classification of numeral image.

The recognition performance of BPNN will highly depends on the network structure. The number of nodes in input, hidden and output layers will determine the network structure. All the neurons of one layer are fully interconnected with all neurons of its just preceding and just succeeding layers (if any). The network consists of 50 nodes in the input layer (corresponding to 50 features). The output layer has 10 neurons corresponding to 10 numerals. Therefore only the number of nodes in the hidden layer is needs to be determined. This architecture has been selected after a considerable experimentation. The architecture of 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 thumb for deciding the number of neurons in the hidden layer.

- The number of hidden neuron should be less than twice the input layer size.
- The number of hidden neuron should be in the range between the size of the input layer and the size of the output layer.
- 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 hidden layer, suitable thumb rule mentioned can be considered individually. For validation purpose 2000 training samples are divided into 1000 samples for learning (for validation) and 1000 samples for testing (for validation). The hidden layer nodes are varied to achieve minimum epochs to converge, optimal training and testing sample performance. We followed these procedures to arrive at 80 neurons for the hidden layer. Again using these 2000 training samples network structure is trained and then 2000 testing samples are shown to the network (not included in training/validation phase) to find the classification and recognition.

Since our desired outputs must be ranged between 0 to 1, so we have selected log sigmoid as the transfer function for both hidden and output layer. We have used 'Mean Squared Error' (MSE) as 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 adopt 'batch gradient descent back propagation' as a learning algorithm.

5.3. Support vector machine for classification and recognition

Support vector machine is new classifier that is extensively used in many pattern recognition applications. SVM uses the principle of structural risk minimization by minimizing Vapnik Chervonenkis (VC) dimensions [13, 14]. On pattern classification problem, SVM demonstrate very good generalization performance in empirical applications.

SVM are binary classifiers that separate linearly any two classes by finding a hyper plane of maximum margin between the two classes. The margin means the minimal distance from the separating hyper plane to the closest data points. SVM learning machines 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 called support vectors.

There are two approaches to extend SVMs for multi-class classification. First one is one against one (ONO) and other is one against all (ONA). We have used ONA approach where N classifiers are performed to separate one of N mutually exclusive classes from all other classes.

An SVM assumes that all samples in the training set are identically distributed and independent. It uses an approximate implementation to the structure risk minimization principle in statistical learning theory, rather than the empirical risk minimization method. A 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. Gaussian kernel performs superior compare to linear kernel, polynomial kernel etc. We have used Gaussian kernel.

5.4. Experimental results and comparative study

In order to evaluate the performance of the proposed method, we consider handwritten Kannada Numerals. We have collected handwriting of 140 individual writers and total of 4000 samples are considered for Kannada numerals. Here 2000 samples are used for training purpose and remaining 2000 samples are used for testing.

For recognition and classification purpose NNC, BPNN and SVM classifiers are used. Table 1 gives the results for Kannada handwritten numeral recognition at different 2000 training and 2000 testing samples. Table 2, Table 3 and Table 4 provides confusion matrix for handwritten Kannada numerals using NNC, BPNN and SVM respectively.

The confusion pair of Kannada handwritten numerals are 3 and 7, which reduces the overall results. This is due to the more similarity between these two numerals. Table 5 provides the comparative results for Kannada handwritten numerals. Figure 4 shows the plot of error (MSE) versus number of iteration for Kannada numeral script.

Table 1.
Kannada handwritten numeral recognition results

Training samples = 2000 and Testing samples = 2000			
Kannada Numerals	NNC classifier	BPNN	SVM classifier
0	99.5	97.5	100
1	96.5	92	98
2	100	94.5	100
3	86.5	83.5	89
4	97.5	95	98
5	88.5	88	91.5
6	94.5	91	94.5
7	93.5	93.5	94.5
8	96	95.5	97
9	97.5	98	98
Average recognition (in %)	**95**	**92.85**	**96.05**

Table 2.
Confusion matrix for Kannada handwritten numerals using NNC classifier

Confusion matrix NNC classifier										
	0	1	2	3	4	5	6	7	8	9
0	**99.5**	0.5								
1	1	**96.5**							2.5	
2			**100**							
3			2	**86.5**	1	1	0.5	8.5	0.5	
4				0.5	**97.5**	0.5			0.5	1
5	1.5	2.5	1	2.5	1.5	**88.5**	0.5	0.5	1	0.5
6							**94.5**	4.5	1	
7				2			4.5	**93.5**		
8	1.5		1.5	0.5				0.5	**96**	
9							1	0.5	1	**97.5**

Table 3.
Confusion matrix for Kannada handwritten numerals using BPNN classifier

Confusion matrix BPNN classifier										
	0	1	2	3	4	5	6	7	8	9
0	**97.5**	1	0.5	0.5			0.5			
1	1.5	**92**		3		1.5			2	
2	0.5		**94.5**	0.5	0.5	4				
3			2	**83.5**	4.5	4.5	1.5	3.5	0.5	
4					**95**	0.5		0.5	3	1
5	1	2		6	2	**88**	0.5	0.5		
6	0.5	0.5		0.5	0.5		**91**	4.5	1.5	1
7				1.5		0.5	3	**93.5**	0.5	1
8	1.5		0.5	1	0.5				**95.5**	1
9				1.5					0.5	**98**

Table 4.
Confusion matrix for Kannada handwritten numerals using SVM classifier

Confusion matrix SVM classifier										
	0	1	2	3	4	5	6	7	8	9
0	**100**									
1	0.5	**98**		1.5						
2			**100**							
3		0.5	1.5	**89**	2	1	1	5		
4					**98**	0.5		0.5	0.5	0.5
5		1.5		3.5	2	**91.5**	0.5	1		
6	0.5	0.5					**94.5**	3.5	1	
7				2			3	**94.5**		0.5
8				2					**97**	1
9							1.5	0.5		**98**

Table 5.
Comparative results for Kannada handwritten numeral

Kannada numerals			
Feature extraction method **Zoning**	Data set size	Classifier	Recognition rate (%)
[10]	2000	NNC	92.4
[10] This paper	4000	NNC	95
[10] This paper	4000	BPNN	92.85
[10] This paper	4000	SVM	**96.05**
[12]	4000	NNC	96.95
[12]	4000	BPNN	94.35
[12]	4000	SVM	97.15

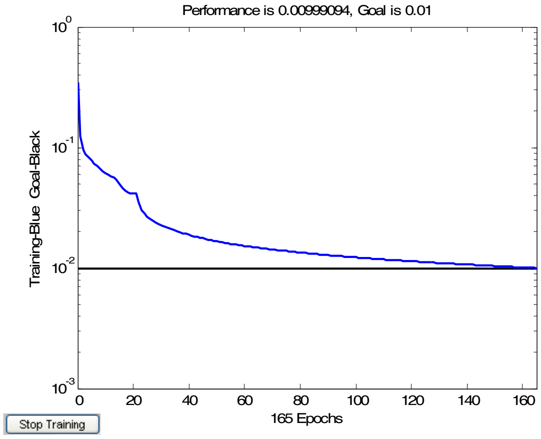

Figure 4. Plot of error versus iterations for handwritten Kannada numerals

6. Conclusion

In this paper we have presented a Image centroid and zone based angle feature extraction system. We have achieved highest recognition rate of 96.05% for Kannada numerals using Support vector machine. Using zone based feature extraction, we have achieved good results even when certain preprocessing steps like filtering, smoothing and slant removing are not considered. Our future work aims to develop new zone based feature extraction algorithms for efficient character classification and recognition and extend our works to other Indian scripts.

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, and 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 three popular South Indian scripts: A novel approach:", *Proceedings of the second International Conference on information processing ICIP*, pp.162-167, 2008.

[9] S.V. Rajashekararadhya, and P. Vanaja Ranjan,"Neural network based handwritten numeral recognition of Kannada and Telugu scripts", TENCON 08 Hyderabad – in press.

[10] S.V. Rajashekararadhya, and P. Vanaja Ranjan,"Efficient handwritten numeral recognition of Kannada and Telugu scripts", *International Conference on Sensors, Security. Software and Intelligent Systems ISSSIS 2009* IP pp.19-23, 2009

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

[12] S.V. Rajashekararadhya, and P. Vanaja Ranjan,"Support vector machine based handwritten numeral recognition of Kannada ", *International conference on advanced computing IEEE IACC 2009*, accepted.

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

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