Kannada, Telugu and Devanagari Handwritten Numeral Recognition with Probabilistic Neural Network: A Script Independent Approach

B.V.Dhandra
Department of P.G. Studies and Research in Computer Science
Gulbarga University, Gulbarga
Karnataka, India.

R.G.Benne
Department of P.G. Studies and Research in Computer Science
Gulbarga University, Gulbarga
Karnataka, India.

Mallikarjun Hangarge
Karnatak Arts, Science and Commerce College, Bidar
Karnataka, India.

ABSTRACT
In this paper a script independent automatic numeral recognition system is proposed. A single algorithm is proposed for recognition of Kannada, Telugu and Devanagari handwritten numerals. In general the number of classes for numeral recognition system for a scripts/language is 10. Here, three scripts are considered for numeral recognition forming 30 classes. In the proposed method 30 classes have been reduced to 18 classes. The global and local structural features like directional density estimation, water reservoirs, maximum profile distances and fill hole density are extracted. A Probabilistic neural network (PNN) classifier is used in the recognition system. The algorithms efficiency is for various radial values of PNN classifiers, with different experimental setup and obtained encouraging results are compared to other methods proposed in the literature survey. A total of 2550 numeral images of Kannada, Telugu and Devanagari scripts are considered for experimentation. The overall accuracy of the system is 97.20%. The novelty of the proposed method is that, it is script independent, thinning free, fast, and without size normalization.

General Terms:
Document Image Analysis

Key words:
OCR, Handwritten Numeral, Indian scripts, Structural feature, Probabilistic Neural Net (PNN)

1. INTRODUCTION
Automatic numeral recognition of a script/language has variety of applications like reading postal zip code, passport number, employee code, bank cheque, form processing and so on. Automatic Numeral recognition is an important component of character recognition system. The problem of the handwritten numeral recognition is a complex task due to the variations among the writers style of writing, shape, stroke etc.

The problem of numeral recognition system is studied for decades and many methods have been proposed such as template matching, dynamic programming, hidden Markov modeling, neural network, expert system and combinations of all these techniques [1, 2]. Feature extraction plays a vital role in image processing system in general and character recognition system in particular. A survey of various feature extraction methods for character recognition can be found in Ivind and Jain [3]. Recognition of character/numeral in foreign languages like English, Chinese, Japanese, and Arabic are reported by many authors. In the Indian context, some major works are reported in Devanagari, Tamil, Bengali and Kannada numeral recognition [4, 5, 6]. Dinesh Acharya *et. al*[7] uses 10-sgment string, water reservoir, horizontal and vertical stroke feature for Kannada numeral recognition, U.Pal *et. al* [8] uses zoning, directional chain code for Kannada numerals recognition. Dhandra *et. al* [9] proposed a method based on directional density feature which is thinning free, independent of size, and font styles of the numeral for printed English numerals. However, all these recognition schemes are script dependent, and their performance good for a languages it is considered. In the Indian context, Sanjeev Kunte and Sudhakar Samuel [12] suggested script independent handwritten Numerals recognition system with wavelet feature and neural network classifier and reported 92.30% average recognition rate.

Thus, from the literature it reveals that there are methods, which suffers from larger computation time mainly due feature extractions for large set and various pre-processing stage, i.e. size normalization, skeletonising or thinning and for neural network training. In addition, the above said recognition systems fail to meet the desired accuracy when they exposed to the different scripts. Hence, it is necessary to develop a method, which is independent of script, and reduces the numeral class used for training; which yields high recognition rate. This has motivated us to design a simple, efficient, and robust algorithm for script independent numerals recognition system.

In this paper, the four different categories of structural features are combined to obtain high degree of accuracy for script independent handwritten numeral recognition system. We have extracted 13 potential feature set includes Directional density estimation [9], Water Reservoir principle based features [10], Maximum profile distances and fill hole density. The proposed method address the extension of Dhandra *et.al* [19] for Script Independent Numeral Recognition based on Structural features.

The paper is organized as follows: Section 2 contains the preprocessing of isolated numerals and gives brief description about the languages considered for recognition and their reduced numeral set. Feature Extraction Method is describes in Section 3. The proposed algorithm is presented in Section 4. The Classifications method is the subject matter of Section 5. The experimental details and results obtained are presented in Section 6. Section 7 contains the conclusion part.

2. NUMERALS PRE-PROCESSING
The standard database for multi script handwritten numeral character is not available; therefore, our own database created for Kannada, Telugu and Devanagari scripts. Data collected from different professionals belonging to schools, colleges, and commercial sectors. The page containing collected multiple lines of isolated handwritten numerals scanned through a flat

bed HP scanner at 300 DPI. The scanned images are binarized using global threshold stored in bmp file format. Scanned isolated numeral images often contain noise that arises due to printer, scanner, print quality, etc. therefore, it is necessary to filter numeral images before we process the numeral recognition. So, the noise removed by using median filter and scanning artefacts are removed by using morphological opening operation

2.1 Language Sets for Recognition

India is a multilingual and multi script country and uses 18 scripts. Hence, there is a need of multilingual and multi-script OCR system for an Indian context. Thus, development of multilingual OCR system in general and the developments of multilingual numeral system in particular are considered as one of the challenging problem to be addressed. Here, we have considered three scripts namely Kannada, Telugu and Devanagari scripts for our experiments as an initial attempt in order to justify the multilingual capability of the proposed system.

Recognition of numeral from multilingual document images is approached by two ways (1) through script identification fallowed by numeral recognition, here identification of the numeral script carried firstly and fallowed by run the recognition of numeral algorithm belongs to script. (2) Recognition of numeral directly without script identification, in this approach total number of classes is increased but it eliminates step of script identification. In this proposed approach, we fallows second method of recognition system.

Table 1 shows English numerals with corresponding Kannada, Telugu and Devanagari numerals; to get an idea about the shape difference between numerals of different scripts.

TABLE 1: NUMERAL SET FOR DIFFERENT LANGUAGE

English	0	1	2	3	4	5	6	7	8	9
Kannada	೦	೧	೨	೩	೪	೫	೬	೭	೮	೯
Telugu	౦	౧	౨	౩	౪	౫	౬	౭	౮	౯
Devanagari	०	१	२	३	४	५	६	७	८	९

From Table 1, by observing, it is clear that some numerals of Kannada, Telugu, and Devanagari have similar structure. For example; numeral '0' has same structure in all three languages Numeral '1','2' of Kannada and Telugu are similar; Numeral '3','4','6' of Telugu and Devanagari are similar. Thus, by using this knowledge, the possible number of classes may be reduced for training phase of recognition system. Hence, the total number of numeral class is reduces to 18 intended classes. The reduced numeral set considered for recognition listed in Table 2. Table 3 contain number of class considered for recognition along with corresponding numeral of three scripts.

TABLE 3: NUMERAL CLASS WITH NUMERAL SCRIPTS AND CORRESPONDING RECOGNIZED NUMERAL

Class Label	Corresponding Numerals	Numeral belongs to
Class 0	K-0,T-0,D-0	Numeral 0
Class 1	K-1,T-1	Numeral 1
Class 3	D-1	
Class 3	K-2,T-2	Numeral 2
Class 4	D-2	
Class 5	K-3	Numeral 3
Class 6	T-3,D-3	
Class 7	K-4,T-4,D-4	Numeral 4
Class 8	K-5,T-5	Numeral 5
Class 9	D-5	
Class 10	K-6,T-6,D-6	Numeral 6
Class 11	K-7	Numeral 7
Class 12	T-7	
Class 13	D-7	
Class 14	K-8,T-8	Numeral 8
Class 15	D-8	
Class 16	K-9,T-9	Numeral 9
Class 17	D-9	

TABLE 2: REDUCED NUMERAL CLASSES FOR RECOGNITION

Languages					Languages				Tot. Class
Eng	Kan	Tel	Dev		Eng	Kan	Dev	Tel	18
0	೦	౦	०		0		०		1
1	೧	౧	१		1	೧		१	2
2	೨	೨	२		2	೨		२	2
3	೩	౩	३	→	3	೩	३		2
4	೪	౪	४		4		४		1
5	೫	౫	५		5		५	५	2
6	೬	౬	६		6		६		1
7	೭	౭	७		7	೭	७	७	3
8	೮	౮	८		8		८	८	2
9	೯	౯	९		9		९	९	2

3. FEATURE EXTRACTION METHOD

Feature extraction is an important component of any recognition system. Selection of feature is probably the single most important factor in achieving high recognition performance. In this paper a structural features are used for the recognition of numerals. The Directional density estimation features, Water Reservoir principle based features, Profile distance features and fill hole density feature are used for the numeral recognition. The normalization of feature vector is carried out by dividing each feature by the maximum value in that vector. All features of test and training images are normalized in the range of (0,1).

3.1 Directional Density Estimation

The outer directional density of pixels is counted row/column wise until it touches the outer border of the character in the four directions viz. left, right, top, and bottom direction as shown in Fig.1. It also exhibits the corresponding directional pixels considered in the count as black band area [9].

a. b. c. d. e

.

Figure 1: Direction Density estimation for Kannada numeral 4

3.2 Water Reservoir Principle Based Features

The water reservoir based principle is as follows. If water is poured from one side of a component, the cavity regions of the component where water will be stored are considered as reservoirs [10].

Top (bottom) Reservoir: The reservoir obtained when water is poured from top (bottom) of the component.

Left (right) Reservoir: When the water is poured from left (right) side of the component, the cavity regions of the components where water will be stored are considered as left (right) reservoir.

Top, bottom, left, and right reservoir of Kannada numerals are illustrated in Figure 2.

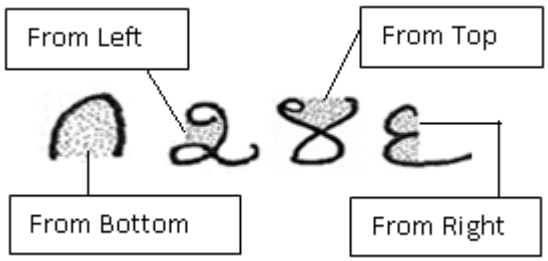

Figure 2: Water Reservoirs in Numerals

3.3 Fill Hole Density

The looping area of the numeral is filled with ON pixels [13], further fill hole density is estimated and taken as a feature.

3.4 Maximum Profile Distances

After fitting the bounding box on each numeral, their profiles are computed in four directions. While computing the profile, we have considered only 40% of the middle area in four directions of the bounding box. Thus the maximum profile is obtained in four directions, the profile feature computations are illustrated in Fig. 3.

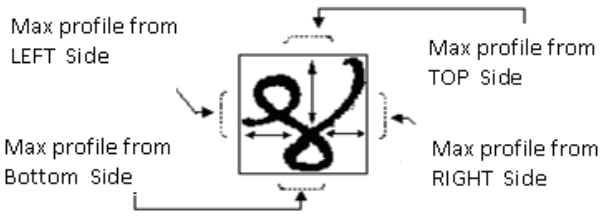

Figure 3: Maximum profile distances from all sides of Bounding Box

4. ALGORITHM

Input : Isolated Binary Numeral of three scripts.
Output: Recognition of the Numeral.
Method: Structural feature and *Probabilistic Neural Network*

Step 1: Preprocess the input image to eliminate the noise and scanning artifacts using median filter and Morphological operator.

Step 2: Fit the minimum rectangle-bounding box for an input image and crop the digit.

Step 3: Extract the structural features and stored in the library.

Step 4: Train the PNN classifier with feature vector stored in the library.

*Step 5:*Classify the test image to its appropriate class label using PNN classifier with various radial value of network.

Step 6: finally, recognize the numeral image as given below
if *test image* € *class 0* → *numeral 0*
else if *test image* € *class 1,2* → *numeral 1*
else if *test image* € *class 3,4* → *numeral 2*
else if *test image* € *class 5,6* → *numeral 3*
else if *test image* € *class 7* → *numeral 4*
else if *test image* € *class 8,9* → *numeral 5*
else if *test image* € *class 10* → *numeral 6*
else if *test image* € *class 11,12,13* → *numeral 7*
else if *test image* € *class 14,15* → *numeral 8*
else *test image* € *class 16,17* → *numeral 9*

Step 7: stop.

The above algorithm should be trained with 1800 numeral images shown in the table 6. After the PNN classifier has been trained on these numeral images, the trained classifier can be used on unknown numeral images.

5. CLASSIFICATION

Probabilistic Neural Network (PNN) classifier: Probabilistic neural network is a kind network suitable for classification problems. The PNN classifier provided a good generalization ability and fast learning capability, which are crucial for handwritten character recognition system. This networks uses a Radial basis transfer function to calculate a layers output from its net input [17]. The architecture of Probabilistic Neural Net is

made up of four units viz. input units, Pattern units[Class A …Class R, total 18 classes], Summation units and Output units as shown in the figure 4.

When an input are presented, the first layer computes distance from input vector to the training vector, and produces a vector whose elements indicate how close the input is to a training input. The second layer sums these contributions for each class of inputs to produce as its net output a vector of probabilistic. Finally, compete transfer function on the output of the second layer picks the maximum of these probabilistic, and produces a 1 for that class and a 0 for the other classes.

In our experiment we use trial and error method to set the spread value. Finally, the experiment is carried out for with selected spread values (viz. 0.1, 0.5, 0.05) for computing distance between the input vectors.

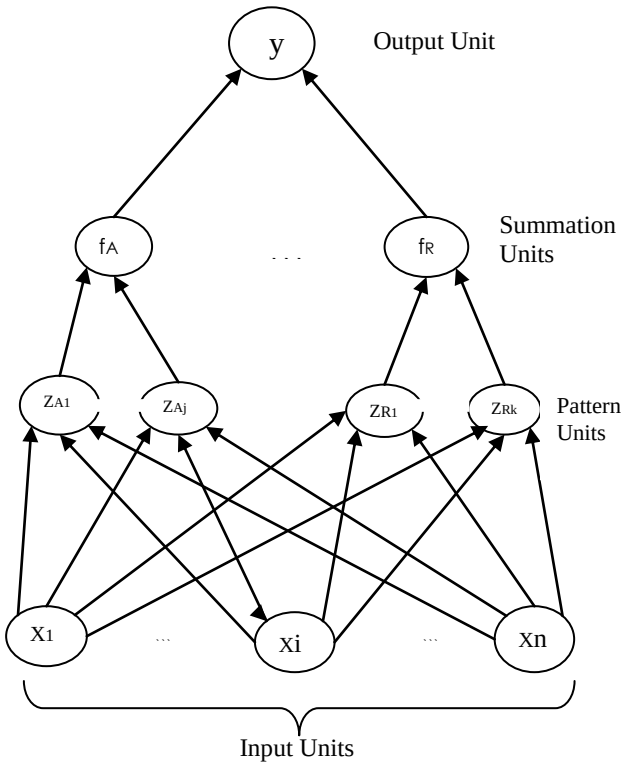

Figure 4: Architecture of Probabilistic Neural Net

6. RESULTS AND DISCUSSION

This algorithm used 2550 handwritten numeral samples for experimentation purpose. Samples of 1800 handwritten numerals is used for training purposes, which includes 500-Kannada, 700-Telugu and 600-Devanagari numeral as shown in the table 4. For, testing 750 numeral images considered, by taking 25 images for each numerals and each script (18x25=750). The average recognition rate for different languages found to be 97.20% as shown in the Table 6. The recognition rate for individual languages are listed in table 5, from table the success rate of Telugu numeral better than Kannada and Devanagari numerals.

TABLE 4: TABLE SHOWS NUMBER OF SAMPLES FOR CLASS AND NUMBER OF CLASSES SELECTED FROM LANGUAGE IS USED FOR TRAINING PURPOSE

Num	Kannada	Telugu	Devanagari
0	100	---	---
1	100	---	100
2	100	---	100
3	100	100	---
4	---	100	---
5	---	100	100
6	---	100	---
7	100	100	100
8	---	100	100
9	---	100	100

TABLE 5: RECOGNITION RATE FOR DIFFERENT LANGUAGES

Languages	Recognition rate %
Kannada	96.80
Telugu	98.80
Devanagari	96.00

TABLE 7: COMPARISON OF RESULTS WITH OTHER METHODS

Method	Features and Classifier used	Data set	% of Acc.
Sanjeev Kunte *et. al*	Wavelet descriptors and Feed forward network classifier. Five script are considered	250 for each numeral	92.30
Previous method	Structural features and k-NN classifier , Three script are consider	2550	92.40
Proposed	Structural features and Probabilistic Neural network (PNN) classifier , Three script are consider	2550	97.20

TABLE 6: RECOGNITION RESULT FOR 1800+750 IMAGE SAMPLES

Lang.	Digits	Train images	Test Images	Correctly Classified	% Correct
KANNADA	0	100	25	25	100
	1	100	25	25	100
	2	100	25	25	100
	3	100	25	24	96
	4	-	25	24	96
	5	-	25	23	92
	6	-	25	24	96
	7	100	25	24	96
	8	-	25	24	96
	9	-	25	24	96
TELUGU	0	-	25	25	100
	1	-	25	25	100
	2	-	25	24	96
	3	100	25	25	100
	4	100	25	25	100
	5	100	25	25	100
	6	100	25	25	100
	7	100	25	24	96
	8	100	25	25	100
	9	100	25	24	96
DEVANAGARI	0	-	25	25	100
	1	100	25	25	100
	2	100	25	24	96
	3	-	25	22	88
	4	-	25	24	96
	5	100	25	25	100
	6	-	25	23	92
	7	100	25	24	96
	8	100	25	24	96
	9	100	25	24	96
Average Recognition rate in %					97.20%

It is difficult to compare results for handwritten numeral recognition with other researchers in the literature, due the differences in experimental settings, methodology, and the size of the database used.

From Table 7, it is clear that, the proposed method gives very good result compare to other methods; proposed by Sanjeev Kunte and Sudhakar Samuel [12] and our previous method.

The above figures shows that the results obtained are very good. However, our algorithm behaves not well under following limitation: The first limitation is that the input of recognition system is only Kannada/Telugu/Devanagari numeral images otherwise wrong recognize. The second limitation is that the input image must be isolated characters.

7. CONCLUSIONS

In this paper, a script independent handwritten numeral recognition system is proposed. The proposed algorithm is script Independent nature, it recognize the Kannada, Telugu, and Devanagari numerals separately or mixed in nature under single algorithm. The average recognition rate is 97.20% for all three languages. In any recognition process, the important steps to address the feature extraction and correct classification methods. The proposed algorithm tries to address both the factors in terms of accuracy and time complexity. The novelty of this method is that, it is script independent handwritten numeral recognition system. This work carried out as an initial attempt for bilingual/ multilingual handwritten numeral recognition system using common algorithm.

8. ACKNOWELDGEMENT

This work is supported by UGC, New Delhi under Major Research Project grant in Science and Technology, (F.No-F33 -64/2007 (SR) dated 28-02-2008). Authors are grateful to UGC for their financial support.

9. REFERENCES

[1] A.L.Koerich, R. Sabourin, C.Y.Suen, "Large off-lineHandwritten Recognition: A survey", Pattern Analysis Application 6, 97-121, 2003.

[2] J.D. Tubes, A note on binary template matching. Pattern Recognition, 22(4):359-365, 1989.

[3] Ivind due trier, anil Jain, torfiinn Taxt, "A feature extraction method for character recognition-A survey ", pattern Recg, vol 29, No 4, pp-641-662, 1996

[4] A.F.R. Rahman, R.Rahman, M.C.Fairhurst, "Recognition of handwritten Bengali Characters: A Novel Multistage Approach", Pattern Recognition, 35,997-1006, 2002.

[5] R. Chandrashekaran, M.Chandrasekaran, Gift Siromaney "Computer Recognition of Tamil, Malayalam and Devanagari characters", Journal of IETE, Vol.30, No.6, 1984.

[6] P.Nagabhushan, S.A.Angadi, B.S.Anami, "A fuzzy statistical approach of Kannada Vowel Recognition based on Invariant Moments", Proc. Of 2nd National Conf. on Document Analysis and Recognition (NCDAR-2003), Mandy, Karnataka, India, pp275-285, 2003.

[7] Dinesh Acharya U, N V Subba Reddy and Krishnamoorthi, "Isolated handwritten Kannada numeral recognition using structural feature and K-means cluster", IISN-2007, pp-125-129.

[8] N. Sharma, U. Pal, F. Kimura, "Recognition of Handwritten Kannada Numerals", ICIT, pp. 133-136, 9th International Conference on Information Technology (ICIT'06), 2006.

[9] B.V. Dhandra, V.S. Mallimath, Mallikargun Hangargi and Ravindra Hegadi, "Multi-font Numeral recognition without Thinning based on Directional Density of pixels", ICDIM-2006,India, pp.157-160, Dec-2006

[10] U Pal and P.P.Roy, "Multi-oriented and curved text lines extraction from Indian documents", IEEE Trans on system, Man and Cybernetics-Part B, vol.34, pp.1667-1684, 2004.

[11] B.V. Dhandra, R.G.Benne and Mallikargun Hangargi, "Handwritten Kannada Numeral recognition based on structural features", IEEE International conference on Computational Intelligence and Multimedia Application", ICCIMA-07, pp.157-160, Dec-2007.

[12] R Sanjeev Kunte and Sudhakar Samuel R.D, "Script Independent Handwritten Numeral recognition".VIE - 2006, pp 94-98, September 2006

[13] R.C.Gonzal, R.E.Woods, "Digital Image Processing", Pearson Education, 2002.

[14] Rajput, G.G., Mallikarjun Hangarge, "Recognition of Isolated Handwritten Kannada Numeral based on Image fusion method", PReMI07,LNCS, Vol. 4815, Springer Kolkatta, pp153-160, 2007.

[15] V. N. Manjunath Aradhya, G. Hemanth Kumar and S. Noushath, Robust Unconstrained Handwritten Digit Recognition Using Radon Transform, Proc. of IEEE-ICSCN 2007, pp-626-629, (2007).

[16] B.V. Dhandra, R.G.Benne and Mallikargun Hangargi, "Isolated Handwritten Kannada Numeral recognition based on Template matching", IEEE-ACVIT -07, pp.1276-1282, Dec-2007.

[17] S.N.Sivanandam, S.Sumathi and S.N.Deepa," Introduction to Neural Networks", The McGraw-Hill publication.

[18] S.V.Rajashekararadhya and P.V.Vanaja Ranjan,"Neural network based handwritten numeral recognition of Kannada and Telugu scripts",TENCON 2008.

[19] B.V. Dhandra, R.G.Benne and Mallikargun Hangargi, "Script Independent Handwritten Numeral Recognition with structural features", ICISP-2009, pp 431-434, Mysore.

[20] B.V.Dhandra, Gururaj Mukarambi, Mallikarjun Hangarge," Zone Based Features for Handwritten and Printed Mixed Kannada Digits Recognition", International Conference on VLSI, COMMUNICATION & INSTRUMENTATION (ICVCI – 2011),April 07th - 09th, 2011, held at Kottayam, Kerala, India, 2011.