Tri-scripts handwritten numeral recognition: a novel approach

Benne R.G.[1], Dhandra B.V.[1] and Mallikarjun Hangarge[2]
[1]P.G. Department of Studies and Research in Computer Science, Gulbarga University, Gulbarga
[2]Department of Computer Science, Karnatak Arts, Science and Commerce College, Bidar,
dhandra_b_v@yahoo.co.in, rgbenne@yahoo.com,mhangarge@yahoo.co.in

Abstract- In this paper an automatic recognition system for isolated handwritten numerals recognition for three popular south Indian scripts. Kannada, Devanagari, and Telugu numeral sets are used for their recognition. The proposed method is thinning free and without size normalization. The structural features viz. directional density of pixels, water reservoirs, maximum profile distancess, and fill hole density are used for handwritten numerals recognition. A Eclidian distance criterion and K-nearest neighbor classifier is used to classify the handwritten numerals. A total of 5250 numeral images are considered for experimentation, and the overall accuracy of 95.40%, 90.20%, and 98.40% for Kannada, Devanagari and Telugu numerals respectively are achived. The novelty of the proposed method is thinning free, fast, and without size normalization.

Keywords- OCR, Handwritten Numeral Recognition, k-NN, Structural feature, and Indian script

Introduction

Automatic numeral recognition has variety of applications in various fields like reading postal zip code, passport number, employee code, bank check, and form processing. 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 like style of writing, shape, stroke etc. The problem of numeral recognition has been 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 [1,2,16]. 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 is presented in Ivind and Jain [3]. Extensive work has been carried out for recognition of character/numeral in foreign languages like English, Chinese, Japanese, and Arabic. 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] use 10-segment string, water reservoir, horizontal and vertical stroke feature for Kannada numeral recognition, U.Pal *et. al* [8] have used zoning and directional chain code for Kannada numerals recognition. Dhandra *et. al*[9] have proposed a method based on directional density feature which is thinning free, independent of size, and font styles of the English numerals. In the Indian context, Sanjeev Kunte and Sudhakar Samuel [12] have suggested script independent handwritten numerals recognition system using wavelet feature and neural network classifier, and reported 92.30% average recognition rate. Though the performance of the said algorithm is better, it is computationally difficult and requires good amount of time to train the numeral images using neural networks. From the literature survey, it is evident that handwritten numerals recognition is still a fascinating area of research to design a robust and efficient Optical Character Recognition (OCR) system in general and handwritten Numeral recognition system in particular. This has motivated us to design a simple and robust algorithm for handwritten numerals recognition system, which is independent of size, slant, ink, and writing style. In this paper, four different categories of structural features are combined to obtain high degree of accuracy for handwritten numeral recognition system. We have identified 13 potential feature sets including four Directional Density feature [9], four Water Reservoir based features [10], four Maximum profile features and a fill-hole density feature. The proposed method addresses the extension of Dhandra *et.al*[11] for Handwritten Numeral Recognition for three Scripts based on structural features. The paper is organized as follows: Section 2 of the paper contains the preprocessing of isolated numerals and gives a brief description of the languages selected for recognition and its data set. Feature Extraction Method is described in Section 3. The proposed algorithm is presented in Section 4. The Classification 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 and further enhancement of the problem.

Data Set and Pre-Processing

The standard database for South Indian numeral script is neither available freely or commercially. Hence, we have created our own numeral database. Data is collected from different professionals belonging to schools, colleges, and commercial sectors. We are sucsessful collecting 2500 unconstrained Kannada numeral samples from 250 writers. Similarly, 1250 Telugu numeral samples from 125 writers and 1500 Devanagari samples from 150 writers. The collected data set containing multiple lines of isolated handwritten numerals are scanned through a flat bed HP scanner at 300 DPI and binarized using global threshold and is stored in bmp file format. The scanned and segmented isolated numeral images quite often contains noise that arises due to printer, scanner, print quality, etc. Therefore, it is necessary to filter those noises before processing the numeral images. The noise is removed by using median filter and scanning artefacts are removed by using morphological opening operation.

Languages 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 and it has potential contribution for scientific and economic advancement of a country. Hence, we have considered three south Indian scripts viz. Kannada, Devanagari, and Telugu for our experiments as an initial attempt.

A sample of Kannada, Telugu, and Devanagari handwritten numerals sets is shown in Fig. 1, Fig. 2 and Fig. 3 respectively and are considered for experiments.

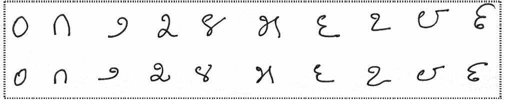


Fig. 1-Sample of Handwritten Kannada numerals

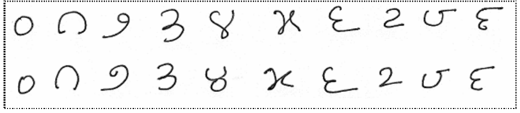


Fig. 2-Sample of Handwritten Telugu numerals

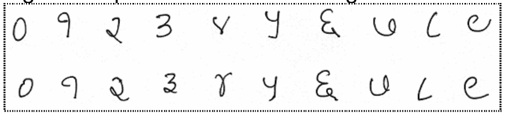


Fig. 3-Sample of Handwritten Devanagari numerals

Feature Extraction Method

Extraction of potential feature is an important component of any recognition system. Selection of potential features is probably the single most important factor in achieving high recognition performance. In this paper, structural features are considered as the potential features and used for the recognition of handwritten numerals. The directional density estimation features, feature based on water reservoir principle, maximum profile distance features, and fill-hole density features are the structural features 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).

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. 4. It also exhibits the corresponding directional pixels considered in the count as black band area [9].

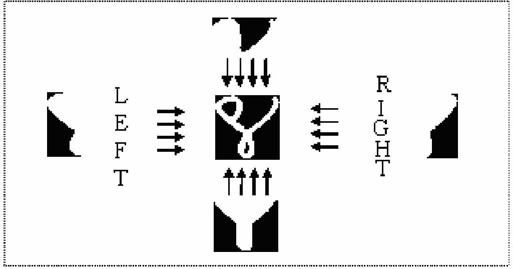


Fig. 4-Directions for Density estimation and pixels consideration

Water Reservoir: The water reservoir based principle is that, 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: The water stored cavity regions of the component, when water is poured from left (right) side of the component will be the left (right) reservoir. Top, bottom, left, and right reservoir of numerals are illustrated in Fig. 5.

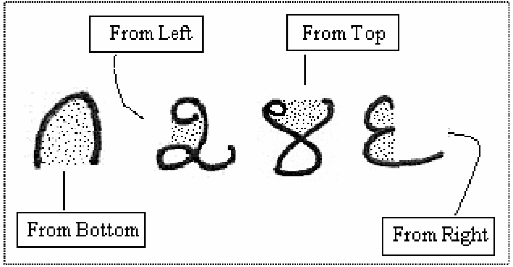


Fig. 5- Water Reservoirs in Numerals

Fill-hole Density: The looping area of the numeral is filled with ON pixels [13] and feature considered is the estimated fill-hole density.

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. 6.

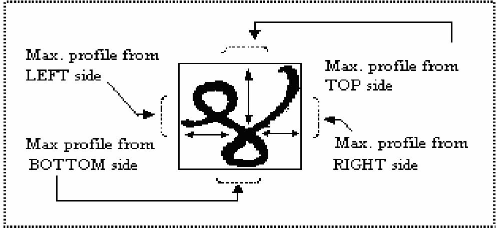

Max. profile from LEFT side
Max. profile from TOP side
Max profile from BOTTOM side
Max. profile from RIGHT side

Fig 6- Maximum profile distances from all sides of Bounding Box

In the following section we exhibit the proposed algorithm for numeral classification.

Algorithm

Input : Isolated Binary Numerals.
Output: Recognition of the Numeral.
Method: Structural features and k-NN classifier.

a. Preprocess the input image to eliminate the noise using median filter and invert the image.
b. Fit the minimum rectangle-bounding box for an input image and crop the digit.
c. Find the directional density of pixels in four direction viz. left, right, top and bottom separately
d. Compute the maximum profile distances from all sides of bounding boxes and fill-hole density.
e. Find the water reservoir in four directions from left, right, top, and bottom.
f. Estimate the distance between feature vector and vector stored in the library.
g. Classify the input image to its appropriate class label using minimum distance K-nearest neighbor classifier.
h. Stop.

Classification

K-Nearest-Neighbor (KNN) classifier: Nearest neighbor classifier is an effective technique for classification problems in which the pattern classes exhibits a reasonably small degree of variability. The k-NN classifier is based on the assumption that the classification of an instance is most similar to the classification of other instances that are nearby in the vector space. It works by calculating the distances between one input pattern with the training patterns. A k-Nearest-Neighbor classifier takes into account only the k nearest prototypes to the input pattern, and the majority of class values of the k neighbors determine the decision. In the k-Nearest neighbor classification, we compute the distance between features of the test sample and the features of every training sample. The class of majority among the k-nearest training samples is based on the minimum distance criteria.

Experimental Results and Discussion

Proposed algorithm uses 2500 handwritten Kannada numerals, 1500 handwritten Devanagari numerals, and 1250 handwritten Telugu numerals for experimentation purpose. The overall accuracy is found to be 95.40%, 90.20%, and 98.40% as shown in the Table 1,2,3 for Kannada, Devanagari and Telugu numerals respectively. The algorithm presents the best performane for Telugu handwritten numerals compared to the Kannada and Devanagari numerals. The average recognition rate of the above three languages is found to be 94.66% and is reasonably high.

Table 1-Kannada Handwritten numeral recognition results

Training samples =2000, Test samples =500 and Number of features = 13		
Kannada Numerals	Testing Samples	Recognition Accuracy
0	50	100.00
1	50	100.00
2	50	98.00
3	50	94.00
4	50	96.00
5	50	90.00
6	50	94.00
7	50	90.00
8	50	94.00
9	50	98.00
Average Recognition		**95.40**

Table 2-Devanagari Handwritten numeral recognition results

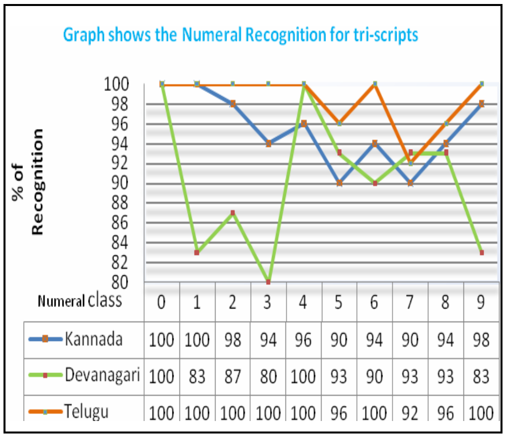
Training samples =1200, Test samples =300 and Number of features = 13		
Devanagari Numerals	Testing Samples	Recognition Accuracy
0	30	100.00
1	30	83.00
2	30	87.00
3	30	80.00
4	30	100.00
5	30	93.00
6	30	90.00
7	30	93.00
8	30	93.00
9	30	83.00
Average Recognition		**90.20**

Table 3-Telugu Handwritten numeral recognition results

Training samples =1000, Test samples= 250 and Number of features = 13		
Telugu Numerals	Testing Samples	Recognition Accuracy
0	25	100.00
1	25	100.00
2	25	100.00
3	25	100.00
4	25	100.00
5	25	96.00
6	25	100.00
7	25	92.00
8	25	96.00
9	25	100.00
Average Recognition		**98.40**

The average recognition rate for individual languages is listed in table 4 and rate recognition of individual numeral class for tri-scripts is as shown in Figure 8. The average accuracy for Devanagari script is brought to 90.20% due to the poor performance rate for recognition of numeral 1,2,3, and 9 otherwise recognition rates is 95%.

The average rate of recognition of Telugu script is very high, but the recognition of numeral 7 for Telugu scripts is falls due to lot of variation of handwritten numeral 7.



Numeral class	0	1	2	3	4	5	6	7	8	9
Kannada	100	100	98	94	96	90	94	90	94	98
Devanagari	100	83	87	80	100	93	90	93	93	83
Telugu	100	100	100	100	100	96	100	92	96	100

Fig. 7- Scatter graph shows the rate of recognition for tri-scripts

Table 4- Recognition rate for different languages

Languages	Recognition rate %
Kannada	95.40
Telugu	98.40
Devanagari	90.20
Average Recognition rate	94.66

It is difficult to compare results for handwritten numeral recognition with those of other methods given in the literature, due to the variations in experimental settings, methodology, and the size of the database used. However, Table 5 presents the comparison of the proposed method with other four methods available for handwritten Kannada numerals.

Table 5-Comparative results for handwritten numerals with other methods

Methods	Features and Classifier used	Data/ feature set	% of Acc.
[8]	Structural features k- means classifier	500	90.50
[14]	Image Fusion Nearest Neighbour	1000	91.20
[15]	Radon transform Nearest Neighbour	1000	91.20
[16]	Template matching, similarity-dissimilarity, binary distance transform, majority voting.	1000	91.00
Proposed Method	Structural features Nearest Neighbour	2500	95.40

From Table 5, it is clear that the proposed method performs better then the other four methods and the recognition rate a 95.40% with less computation cost and with the basic classifier. However, the higher recognition rate can achieved by way of using better classifiers and/or adding new feature set and/or improving preprocessing steps like smoothing, filtering, slant correction.

Conclusion

In this paper, three South Indian popular scripts are consider for recognition of handwritten numerals. The proposed handwritten numeral recognition system uses 13 structural features and a K-NN classifier. The average recognition rate is 94.66% for all the three languages. In any recognition process, the important steps are to address the feature extraction and correct classification method. 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 thinning free, fast, and without size normalization. This work is carried out as an initial attempt for bilingual/multilingual handwritten numerals recognition system.

ACKNOWLEDGEMENT

This work is part of the UGC Major Research Project sanction to the first and third authors (F.No-F33-64/2007(SR) dt. 28-02-2008) in Science and Technology. Authors are grateful to Dr. P. Nagabhushan, Dr. G. Hemanth and Dr. D. S. Guru of University of Mysore for their discussion and critical comments.

References

[1] Koerich A.L., Sabourin R., Suen C.Y. (2003) *Pattern Analysis Application* 6, 97-121.

[2] Tubes J.D. (1989) *Pattern Recognition*, 22(4):359-365.

[3] Ivind due trier, anil Jain, torfiinn Taxt (1996) *Pattern Recg*, 29 (4), 641-662.

[4] Rahman A.F.R., Rahman R., Fairhurst M.C. (2002) *Pattern Recognition*, 35,997-1006.

[5] Chandrashekaran R., Chandrasekaran M. (1984) *Journal of IETE*, 30, 6.

[6] Nagabhushan P., Angadi S.A., Anami B.S. (2003) *Proc. Of 2nd National Conf. on Document Analysis and Recognition (NCDAR-2003), Mandy, Karnataka, India,* pp275-285.

[7] Dinesh Acharya U., Subba Reddy N. V. and Krishnamoorthi (2007) *IISN-2007*, pp-125-129.

[8] Sharma N., Pal U., Kimura F. (2006) *9th International Conference on Information Technology (ICIT'06),* 133-136.

[9] Dhandra B.V., Mallimath V.S., Mallikargun Hangargi and Ravindra Hegadi (2006) *IEEE International conference on Digital Information Management (ICDIM-2006) Bangalore, India*, 157-160.

[10] Pal U. and Roy P.P. (2004*) IEEE Trans on system, Man and Cybernetics-Part B*, 34, 1667-1684.

[11] Dhandra B.V., Benne R.G. and Mallikargun Hangargi (2007) *IEEE International conference on Computational Intelligence and Multimedia Application", ICCIMA-07*, 157-160.

[12] Sanjeev Kunte R. and Sudhakar Samuel R.D. (2006) *VIE-2006,* 94-98.

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

[14] Rajput G.G., Mallikarjun Hangarge (2007) *PReMI07,LNCS, 4815, Springer Kolkatta,* 153-160.

[15] Manjunath Aradhya V. N., Hemanth Kumar G. and Noushath S. (2007) *Proc. of IEEE-ICSCN 2007*, pp-626-629.

[16] Dhandra B.V., Benne R.G. and Mallikargun Hangargi (2007) *IEEE-ACVIT-07*, 1276-1282.