IJMI
International Journal of Machine Intelligence
ISSN: 0975-2927 & E-ISSN: 0975-9166, Volume 3, Issue 3, 2011, pp-121-125
Available online at http://www.bioinfo.in/contents.php?id=31

PRINTED AND HANDWRITTEN KANNADA NUMERALS RECOGNITIONUSING DIRECTIONAL STROKE AND DIRECTIONAL DENSITY WITH KNN

DHANDRA B.V.[1], BENNE R.G.[2], MALLIKARJUN HANGARGE [*]
[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
*Department of Computer Science, Governament First Grade College, Bhalki
*Corresponding author. E-mail: rgbenne@yahoo.com

Received: September 29, 2011; Accepted: November 03, 2011

Abstract- In real life applications number of document contains printed as well as handwritten numerals in a single document. The process of recognition of such mixed numeral with respect to single OCR is complicated task. In this paper, we present a novel method for recognition of printed and handwritten Isolated Kannada numerals using single OCR system. We considered Directional Stroke and Directional density based feature for recognition system is proposed. The proposed system extracts Directional stroke on various angles and Directional density/profile on four side of the numeral image. Further, the stroke and density based extracted feature is feed for recognition systems. A Euclidian distance criteria and k-NN classifier is employed to classify the numeral class. A total 5000 numeral images, which includes 4000 for handwritten images and 1000 for printed images considered for experiments and overall accuracy found to 98.04%. The novelty of the proposed method is thinning free, and without size normalization.
Keywords- OCR, PNN, Structural feature, Handwritten Numeral Recognition, and Indian script

1. Introduction

Optical Character Recognition (OCR) is one of the most widely researched areas in pattern recognition applications. The works on this topic mainly focus on recognizing high quality printed text documents. On the other hand, recognition of handwritten characters has received less attention. However, recognition of handwritten and printed documents is becoming a very important pattern recognition problem updating of human needs. Handwritten and printed numeral recognition is an integral part of any character recognition system. Such types of system is popular due to its variety of applications in various fields like reading postal zip code, passport number, employee code, bank check, and form processing. 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., and printed numeral due various font sizes and font styles.

The researchers for character recognition have proposed various approaches, but most of them are attempt for non-Indian character recognition. The character recognition system of foreign languages like English, Chinese, Japanese, and Arabic is reach to almost success rate. In the Indian context, some major works reported in Devanagari, Tamil, Bengali and Kannada character recognition [4, 5-6]. The problem of character recognition has studied for decades and many methods have proposed such as template matching, dynamic programming, hidden Markov modeling, neural network, expert system and combinations of all these techniques [1-2, 16]. Any character recognition system, feature extraction plays a vital role and selection of appropriate features is most important factor; it helps to increase the success rate. Ivind and Jain [3] present a survey of various feature extraction methods for character recognition.

The various authors attempted to recognize handwritten Kannada numerals with various techniques. Dinesh Acharya [7] have use 10-segment string concept, water reservoir, horizontal and vertical stroke, and end point feature with k-means cluster to Kannada numeral recognition, U.Pal [8] have used zoning and directional chain code for Kannada numerals recognition. Dhandra [11,16] have proposed a method based on directional density feature which is thinning free, independent of size, and font styles of the English numeral and they also proposed Template method with various similarity /dissimilarity measures used for handwritten Kannada numeral recognition system. The recognition system of printed and handwritten for Kannada/Marathi numerals using Fourier descriptor with SVM classifier have been proposed by Rajput [14,19] The concept of Script independent handwritten numerals recognition system is suggested by Dhandra [12] using wavelet feature and structural feature. Ramtake [18] an attempt made to

recognize the Marathi handwritten numeral using invariant moments. The recognition of handwritten numerals for three Indian script is proposed by Dhandra [21].

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 mixed numeral recognition system in particular. This has motivated us to design a simple and robust algorithm for handwritten and printed numerals recognition system under common algorithm; which is independent of size,

2. Data set and pre-Processing

India is a multilingual and multi script country and uses 18 scripts. Kannada language is one of the most popular south Indian scripts. The Kannada script consists of 16vowels, 36 consonants and 10 numerals. The number of documents contains printed as well as handwritten characters as shown in Fig. (1). The process of such document is difficult. Hence, there is a need of Kannada OCR system, which recognizes printed and handwritten character for an Indian context. Thus, development of printed and handwritten Kannada OCR system in general and the developments of mixed numeral recognition system in particular considered as one of the challenging problem, to be address here. Hence, we have considered printed and handwritten numerals for our experiments as an initial attempt.

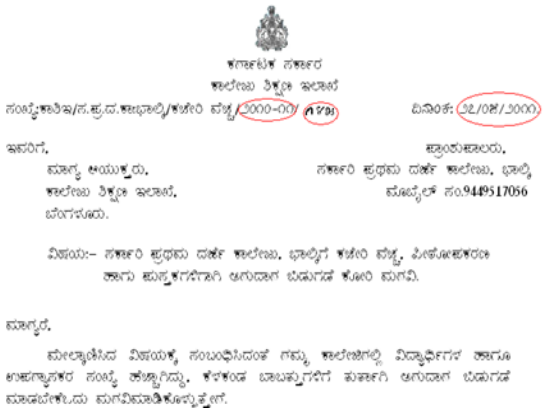

Fig. 1 Sample document containing handwritten and printed Kannada numerals

The printed and handwritten standard database for South Indian numeral script is neither available freely or commercially. Hence, we have created our own printed as well as handwritten numeral database.

Handwritten data collected from different professionals belonging to schools, colleges, and commercial sectors. We are successful collecting 4000 unconstrained handwritten Kannada numeral samples from 200 writers. Fig.(2) shows sample of handwritten Kannada numerals.

Printed data collected from Nudi-4.0 and Baraha-8.0 software. The printed numerals come in multi font and font, and writing style. In this paper, different categories of directional stoke and density based features combined to obtain high degree of recognition accuracy for Kannada numerals are proposed.

The paper organized as follows: Section 2 of the paper contains the preprocessing of isolated numerals and data set for selected languages. Feature Extraction Method described in Section 3. The Classification method and algorithm is the subject matter of Section 4. The experimental details and obtained results presented in Section 5. Section 6 contains the conclusion part and further enhancement of the problem.

multi sizes. We collect 10 numerals of different font sizes from 16 to 50 and various font-styles like, BRH-Kannada, BRH-Amerikannada, BRH-Kailasm, BRH-Vijay, BRH-kasturi, BRH-Bangaluru, BRH-Sirigannada, BRH-Kannada Extra, KGP_kbd, Nudi Akshara-01, Nudi Akshara-02, Nudi Akshara-03, NudiAkshara- 04, Nudi Akshara-05, Nudi Akshara-06, Nudi Akshara-07, Nudi Akshara-08, Nudi Akshara-09, Nudi B-Akshara, and NudiAkshara. We are successful collecting 1000 unconstrained printed Kannada numeral samples. Fig. (3) shows various sample printed Kannada numerals with corresponding font styles.

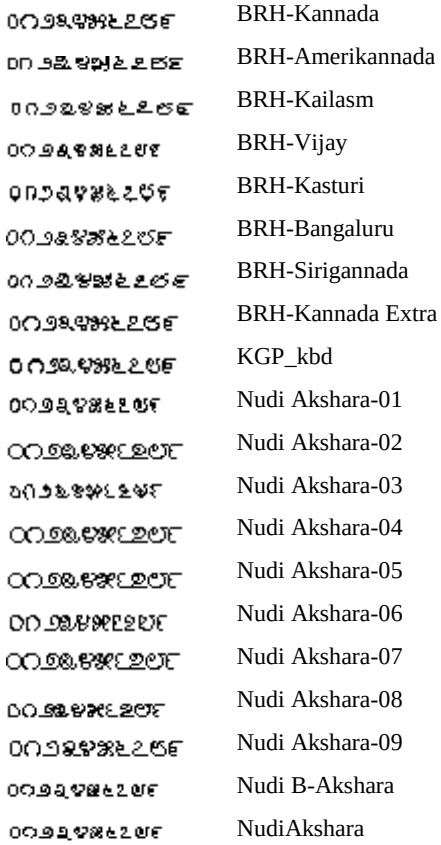

Fig. 2 Printed page contains sample of handwritten Kannada numerals

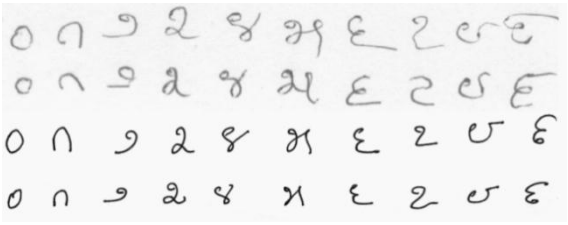

Fig. 3 Printed Kannada numerals with various font styles.

Printed and handwritten collected data set [printed page] containing multiple lines of isolated 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 document segmented automatically by using morphological operators. The segmented isolated numeral images quite often contain 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 removed by using median filter and scanning artefacts are removed by using morphological opening operation.

3. 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 considered as the potential features they are directional stroke and directional density. The list of proposed feature based on directional stroke and directional profile density described below.

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. Figure 1 shows direction density estimation of four sides for Kannada numeral 4. We considered directional pixels in the count as black band area only for all sides of numeral as shown below figure.

Figure 4: Direction Density estimation for Kannada numeral 4

1. Directional Density pixels of an image with respect to image size is computed for four side of image(1x4 feature vector)

$$DD1 = \sum_{i=1}^{N} \frac{Onpixel(\text{Pattern i})}{\sqrt{size}(\text{Pattern i})} \quad [1]$$

Where j= 1…4 belongs left, right, top, and bottom

2. Directional Density pixels of an image with respect to off pixels is computed for four side of image(1x4 feature vector)

$$DD2 = \sum_{i=1}^{N} \frac{Onpixel\ (\text{Pattern i})}{Offpixel\ (\text{Pattern i})} \quad [2]$$

3.1 Directional Stoke estimation: the directional stroke of an image computed on particular direction using morphological transformation with line structuring element. Directional stroke based feature namely; Stroke density, stroke length [maximum] and number of stroke are computed for an angle Θ = 0, 30, 60, 90,120 and, 150.Figure 5shows the stroke effect on various angles for Kannada numeral 3 and 4. The stroke density and largest stoke length is computed by using equation 3 and 4.

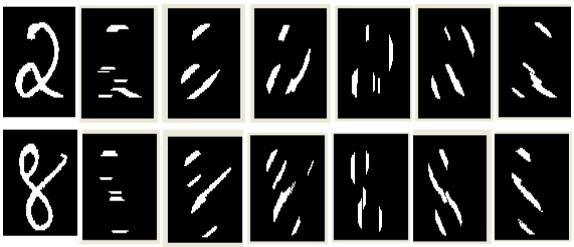

Figure 5: Direction stroke for Kannada numerals 3 and 4 for an angle Θ = 0, 30, 60, 90,120 and, 150.

$$\boldsymbol{DS1} = \frac{\sum Line\ at\ \emptyset\ f^1(i,j)}{\#\ Object - pixel}\ where\ \emptyset \in 0^0, 30^0, 60^0, 90^0, 120^0, 150^0 \quad [3]$$

$$\boldsymbol{DS2} = \frac{Largest\ line}{\#\ Object - pixel} \quad [4]$$

The normalization of feature vector is carries 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) to became a size independent approach.

4. 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 based on, the assumption that the classification of an instance is most similar to the classification of other instances that is nearby in the vector space. It works by calculating the distances between one input patterns 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 Euclidian minimum distance criteria.

Algorithm

Input: Isolated mixed Kannada numeral.
Output: Recognition of the numerals.
Method: Directional stroke & density feature and KNN

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 Directional stroke & density based feature and stored in the library.
Step 4: Normalize the feature vector

*Stept5:*Classify the test image to its appropriate class label using KNN classifier with Euclidian minimum distance criteria.
Step 6: Stop.

5. EXPERIMENTAL RESULTS AND DISCUSSION

Proposed algorithm uses 4000 handwritten and 1000 printed Kannada numerals for experimentation purpose.The experiments carried out separately for printed data set, handwritten data set and mixed numeral data set. The handwritten data set and mixed numeral is divided into 4 folders: 3 folder used for training and 1 folder for testing purpose, whereas 1 folder for training and I folder for testing used in case of printed numeral recognition. The overall accuracy found to be 97.90%, for handwritten numeral, 99.40% for printed numeral and 98.04 for mixed numeral recognition as shown in the Table 1 to Table 3 for Kannada numerals.

Table 1-Kannada Handwritten numeral recognition results

Training samples =3000, Test samples =1000 KNN=3		
Kannada Numerals	Testing Samples	Recognition Accuracy
0	100	100.00
1	100	100.00
2	100	99.00
3	100	97.00
4	100	98.00
5	100	97.00
6	100	97.00
7	100	96.00
8	100	97.00
9	100	98.00
Average Recognition		**97.90**

Table 2- Kannada Printed numeral recognition results

Training samples =500, Test samples =500 KNN=3		
Kannada Numerals	Testing Samples	Recognition Accuracy
0	50	100.00
1	50	100.00
2	50	100.00
3	50	98.00
4	50	100.00
5	50	100.00
6	50	100.00
7	50	98.00
8	50	98.00
9	50	100.00
Average Recognition		**99.40**

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 4 presents the comparison of the proposed method with other methods available for handwritten, printed, and mixed numerals.

Table 3- Kannada mixed numeral recognition results

Training samples =3750, Test samples =1250 KNN=3		
Kannada Numerals	Testing Samples	Recognition Accuracy
0	125	100.00
1	125	100.00
2	125	99.20
3	125	96.80
4	125	97.60
5	125	96.00
6	125	96.00
7	125	95.20
8	125	97.60
9	125	98.40
Average Recognition		**98.04**

Table 4:

Comparative results for numerals with other methods

Methods	Features and Classifier used	Data set	% of Acc.
Results for handwritten numerals for other methods			
[8]	Structural features, k- means classifier	500	90.50
[14]	Image Fusion method , 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
[11]	Structural features with Nearest Neighbour	2500	95.40
Results for Printed numerals for other methods			
[9]	Directional density estimation and Nearest Neighbor	1000	100
[17]	Water Reservoir, Directional density, Max-profile and KNN	1150	100
Results for Mixed numerals for other methods			
[19]	Chain code feature and SVM	2500	97.76
proposed	Directional density, directional stroke features and KNN	5000	98.04

From Table 4, it is clear that the proposed method gives high recognition accuracy by using few simple Directional stoke, Directional density based features, and K-NN classifier. Proposed method gives acceptable accuracy compare to other methods found in the literature survey to the best of my knowledge.

6. Conclusion

In this paper, handwritten Kannada numerals recognition system is proposed. The proposed system uses Stroke and Density features and a K-NN classifier. The average recognition rate of mixed Kannada numeral is 98.04%. In any recognition process, the important steps 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, free from zoning and without size normalization. This work, carried towards an attempt for bilingual/multilingual handwritten numerals recognition system.

References

[1] Koerich A.L., SabourinR. and SuenC.Y. (2003)*Pattern Analysis Application,* 97-121.

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

[3] Ivind due trier, Anil Jainand TorfiinnTaxt, (1996)*Pattern Recognition, 29(4),*641-662.

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

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

[6] Nagabhushan P., Angadi S.A. and Anami B.S. (2003) *NCDAR-2003, Mandy*, India, 275-285.

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

[8] Sharma N, Pal U. and Kimura F.(2006) *ICIT-2006,* 133-136.

[9] Dhandra B.V., Mallimath V.S., Mallikargun Hangargi and RavindraHegadi(2006) *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 Mallikarjun Hangargi(2007) *International conference on Multimedia and Application(IEEE-ICCIMA-07),* 157-160.

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

[13] Gonzal R.C. and Woods R.E. (Edn-2002), *Digital Image Processing.*

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

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

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

[17] Dhandra B.V., Benne R.G.and Mallikargun Hangargi (2007) *National conference on eIT-2007, Baramati(M.S.),*193-199.

[18] Ramteke R.J. and Mehrotra S.C. (2008) *International journal of Computer processing of Oriental languages*.

[19] Rajput G.G., Rajeswari Horkeri, Sidramappa C.(2010)*International Journal of Computer Science and Engineering,* 02,1622-1626.

[20] Dhandra B.V., Benne R.G. and Mallikargun Hangargi(2009) *International Journal for Advances in Computational Research,* 1(2), 47-51.