Mixture of Printed and Handwritten Kannada Numeral Recognition Using Normalized Chain Code and Wavelet Transform

Shashikala Parameshwarappa[1], B.V.Dhandra[2]

[1]Department of Computer Science and Engineering College, Govt.Engg.College, Raichur,Karnataka,India

[2]Department of P.G. Studies and Research in Computer Science, Gulbarga University Kalaburagi, Karnataka,India.

Abstract: ***Optical character recognition (OCR) is one of the most successful applications of automatic pattern recognition. The current research in OCR is now addressing documents that are not well handled by the available systems, including severely degraded, omnifont machine-printed text and (unconstrained) handwritten text. In this paper, a novel method for recognition of printed and handwritten mixed isolated Kannada numeral is presented. An algorithm is proposed to recognize the printed and handwritten Kannada numerals based on shape features such as normalized chain codes and wavelet filters. A normalized chain code and two-dimensional discrete wavelet transforms are proposed to extract as a feature vector of size 22 from the normalized binary images of size 64x64. The SVM and KNN classifier with 2 fold cross validation is applied for classification of handwritten and printed mixed Numerals. The proposed algorithm is experimented on a data set of 4000 numeral images consisting of handwritten and printed numerals. Further the proposed system achieves the average recognition accuracy of 98.04% and 99.12% for mixed Numerals by KNN and SVM classifiers respectively. It achieves reasonably high recognition accuracy with less number of features set.***

Keywords: Kannada Numerals, OCR, Normalized chain code, wavelet transform, SVM classifier.

1. Introduction

Optical character recognition (OCR) is a process of automatic computer recognition of characters in optically Scanned and digitized pages of text. OCR is one of the most fascinating and challenging areas of pattern recognition with various practical application potentials. It can contribute immensely to the advancement of an automation process and can improve the interface between man and machine in many applications. Some practical application potentials of OCR system are: (1) reading aid for the blind, (2) automatic text entry into the computer for desktop publication, library cataloging, ledgering, etc. (3) automatic reading for sorting of postal mail, bank cheques and other documents, (4) document data compression: from document image to ASCII format, (5) language processing, (6) multi-media system design, etc.

The recognition of machine printed and handwritten digits has been the subject of much attention in pattern recognition because of its number of application such as mail processing, automatic data entry, bank check reading, reading of the customer filled forms, vehicle registration numbers and many more. Advancement of e-technology has made the revolution on every field in general and document automation in particular. Printed and handwritten numerals may appear together in documents like application forms, postal mail. This revolution made to develop an OCR system for different languages and scripts for printed and hand printed documents to process automatically. In this direction, many researchers have developed the numeral recognition systems by using various feature extraction techniques such as global transformation and series expansion features like Hough Transform, Fourier transform, statistical features and geometrical features. Extensive work has been carried for recognition of characters in foreign languages like English, Chinese, Arabic. Lot of contribution can be found for printed characters compared to handwritten characters. A Brief amount of literature review is explained below.

U. Pal et al. [1] have used zoning and directional chain code as features vector of size 100 for handwritten Kannada numeral recognition, and obtained reasonably high recognition accuracy. N. Sharma et al. [2] have considered the directional chain code information of the contour points of the characters as features set for recognition of the handwritten Devanagari numerals and characters by using quadratic classifier. Their recognition accuracy is 98.86% and 80.36% respectively. S.V. Rajashekararadhya et al. [3] have proposed zone centriod and image centriod based angle feature extraction system for isolated Kannada numerals recognition and reported 97.3% accuracy. Dhandra et al. [4] have proposed spatial features as feature vector of size 13 for handwritten Kannada numerals and vowels recognition have reported overall accuracy of 96.2% and 90.1%. G. Raju et al. [5] have proposed wavelet packet as a feature set for recognition of handwritten Malayalam (one of the south Indian languages) characters. Feed forward neural network architecture is used for classification and obtained the recognition accuracy of 90%. Rajput et al. [6] have proposed Fourier descriptors and Zone based chain code features amounting to 608 features for handwritten Kannada numerals and vowels recognition and achieved the recognition accuracy of 98.45% and 93.92% respectively. Their accuracy is reasonably high but at the same time complexity of the algorithm is large due to large feature set. Sanjeev Kunte et al. [7] have proposed an OCR system for the recognition of basic characters of printed Kannada text, which works for different font size and font style. Each image is characterized by using Hu’s invariant and Zernike moments. They have achieved the recognition accuracy of 96.8% with Neural Network classifier. Rajput et al. [8] have proposed chain code and fourier code descriptors features for printed and handwritten numerals recognition and have reported the overall recognition accuracy of 97.76%. From the literature survey, it is evident that still handwritten character/numeral recognition is a fascinating area of research to design a single optical character recognition (OCR) system.

Hence in the proposed study an attempt is made to use the Normalized chain code and wavelet based features for the recognition of mixture of handwritten and printed Kannada numerals.

The Section 2 of this paper provides information about data collection and pre-processing methods. The Section 3 deals with the feature extraction method and designing of the proposed algorithm for printed and hand printed Kannada numerals character recognition system. The Experimental results and discussion are discussed in Section 4 and the Section 5 concludes the paper.

2. Data Collection and Preprocessing

Optical character recognition refers to the branch of computer science that involves reading text from paper and translating the images into a form that computers can manipulate (like ASCII codes, for instance). It is observed that, to validate and verify the results of the proposed algorithm the standard database for handwritten Kannada character and printed Kannada script are not available in the state of the art literature. Hence, we build our own database for the purpose of experimentation and validation. Totally 2000 Kannada hand printed numeral images are collected from the varies professionals belonging to Primary Schools, High Schools and Colleges. The printed datasets are modeled by using Nudi and Baraha software's with 7 different font and styles. The printed dataset contains multi-font and multi-size numerals. The collected handwritten and printed Kannada document contains a multiple lines of Kannada scripts. These documents are scanned through a flatbed HP scanner at 300 dpi which usually yields a low noise and good quality document image were cropped up manually and stored as gray scale images. Binarization of image is performed using Otsu's global thresholding method and is stored in bmp file format. The raw input of the digitizer typically contains noise due to erratic hand movements and inaccuracies in digitization of the actual input. The noise present in the image is removed by applying median filter. A minimum bounding box is then fitted to the numeral images. To bring uniformity among the cropped image is normalized to 64x64 pixels. The Fig. 1 and Fig. 2 shows the sample dataset of handwritten and printed Kannada numerals respectively.

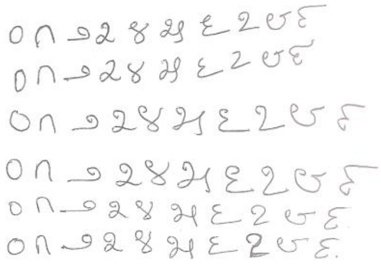

Figure 1: Handwritten Kannada Numerals

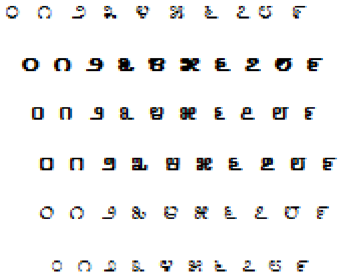

Figure 2: Printed Kannada Numerals

3. Feature Extraction

The feature extraction is described about the characteristics of an image. It is one of the most important components for any recognition system, since the classification/recognition accuracy is depending on the features. Two well known feature extraction methods 1&2. Those are Normalized chain code and wavelet decomposition for printed and handwritten Kannada Numerals recognition system is proposed. A brief description about Normalized chain code and wavelet decomposition is given below.

3.1. Chain Code

Chain codes is used to represent the boundary of an object composed of pixels of regular cells by connected sequence of straight-line segments of specified length and direction. The object is traversed in clockwise manner. Chain codes are one of the shape representations which are used to represent a boundary is based on 4-connectivity or 8-connectivity of the segments [9]. The direction of each segment is coded by using a numbering scheme as shown in Fig 3. According to [10], chain code can be generated by a boundary of an object in a anticlockwise direction and assigning a direction to the segments connecting every pair of pixels.

Chain Code method's work as follows:

Start:

1. Choose a starting pixel anywhere from the object boundary.There must be an adjoining boundary pixel at one of the eight locations surrounding the current boundary pixel as shown in Figure 3.
2. If the pixel found is located at the right of the current location or pixel, a code "0" is assigned.
3. If the pixel found is directly to the upper at the right, a code "1" is assigned. Similarly code "2" & code "3" is assigned as shown in Fig.3a.

4. The process of locating the next boundary pixel and assigning a code is repeated until we came back to our first location.

End.

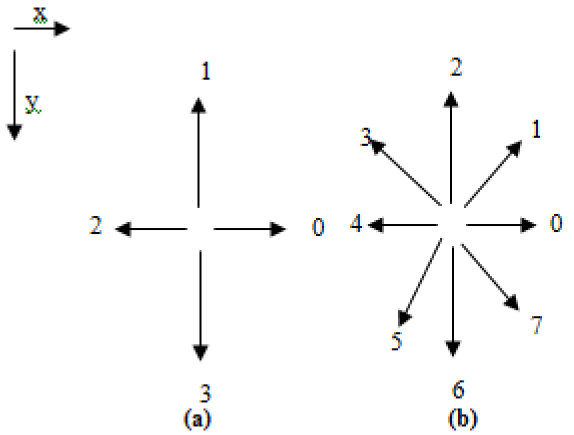

Figure 3: Direction numbers for (a) 4-directional chain codes, (b) 8-directional chain code

The process of finding the boundary pixel and assigning a code is shown in Fig 4.

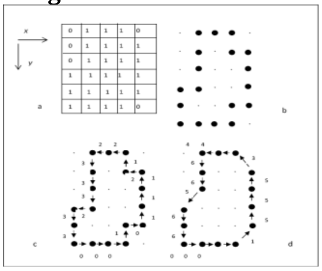

Fig. 4 (a) & (b) A 4-connected object and its boundary; c & d) Obtaining the chain code from the object in (a & b) with (c) for 4- connected and (d) for 8-connected

Chains are used for description of object borders.

Symbols in a chain usually correspond to the neighborhood of primitives in the image.

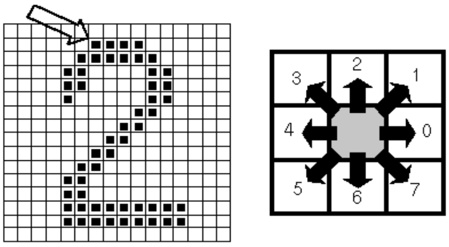

Figure 3.1 An example chain code; the reference pixel is marked by an arrow: 0000776655555556600000006444444442221111112234445652211.

It is observed that the chain code for different Numerals has different length code and length of each chain code depends on the size of the character image. More ever length of chain code is very high in case of certain Numeral image. We have solved this problem by normalizing the chain code values as explained below. The following chain code is generated for Numerals by traversing it in anticlockwise direction.

V1= [1 2 2 0 0 0 0 0 0 0 6 7 6 6 7 6 6 6 6 6 7 6 6 6 6 5 6 6 6 6 5 5 6 5 4 4 5 4 5 4 5 4 4 4 5 4 4 4 4 3 4 4 4 3 4 4 2 3 3 4 3 4 4 3 2 3 3 2 3 1 1 1 1 4 4 4 6 6 6 2 2 2 2 4 4 2 3 2 3 2 2 2 2 2 2 2 2 2 1 1 1 4 7 7 7 3 3 2 2 1 1 5 5 5]

The frequency of the codes 0, 1, 2…..7. for vector *V1* is given below in V2 .

V2= [7 10 23 13 25 10 21 6].

The normalized frequency, represented by vector *V3*, is computed by using the formula

V3= *V2* / │ *V1*│ where │ *V1*│= ∑ *V2*

The resultant normalized frequency of the chain code for V3 is

V3= [0.06087 0.08696 0.20000 0.11304 0.21739 0.08696 0.18261 0.05217]

Hence, the desired feature vector of size 08 is V3. The algorithm for computing normalized chain codes is given section 3.3

3.2. Wavelet Decomposition.

Wavelets transform results in sub images corresponding to smooth component in the three directions horizontal ,vertical and diagonal .The information content of smooth component and high pass filtered components in the three directions should unique feature of an image. This feature can be characterized with number of zero crossing of wavelet coefficient in each sub image. The count of zero-crossing in all sub images together shall be used as a character feature. The algorithm for computing Feature extraction is given below.

Training Phase:

3.3 Algorithm: Feature Extraction

Method:

Input: preprocessed Isolated Kannada Numerals image.

Output: Feature library.

Step 1.

Start:

1.1 Trace the boundary in counterclockwise Direction and generate 8 dimensional chain codes 0 to 7 (Fig.4).

1.2 Compute the frequency of the codes 0 to 7.

1.3 Divide frequency of each code by sum of the frequencies. To obtain feature vector of length 08.

1.4 Repeat the step 1.1 to step 1.3 for all sample images representing Numerals considered for training.

Step 2.

2.1 Apply two level forward wavelet packet transform using db4 filter.

2.2 For each sub band count the number of zero crossing row wise,columnwise,daignalwise.

2.3 Obtain the feature vector of length 14.

2.4 Repeat the step 2.1 to step 2.3 for all sample images representing numerls considered for training.

Step 3.

Store the computed normalized chain code from step 1 and wavelet decomposition coefficient from step 2 as the features stored in train library in the database.

End.

The eight features of chain code and fourteen features of wavelet decomposition coefficients are given to the KNN and SVM classifier for classification.

Testing Phase:

3.4 Algorithm: Recognition of Kannada Numerals

Input: Isolated test Numeral image.
Output: Recognition of the Numeral.
Start:
Step l**:** Extract the features by using Algorithm 3.3.
Step 2**:** Compute the distance between the feature vectors of the test image and with the feature vector of the trained image stored in knowledge base.
Step 3: Minimum distance computed in the above step 2 is the recognized Numeral.
End.

4. Experimental Results and Discussion

The proposed method is implemented using Intel Core 2 Quad processor @ 2.66 GHz machine and MATLAB 2012b. Based on the KNN and SVM classifiers 2000 Kannada handwritten and 2000 Kannada printed digits are classified. The data set for handwritten numerals consists of 2000 images and with 200 images representing each class/numerals, considering 50% for training and 50% for testing. For result computation we have used k-fold cross validation technique. The experiments were carried out by varying the values of k i. e .k= 1, 3, 5 and found optimal result when k =3 as shown in Table 3.The average percentage of recognition accuracy is 96.61 and 97.88 was obtained for handwritten numerals.

The experimental results are obtained by using 2000 training samples and 2000 testing samples for mixed Kannada handwritten and printed digits (Table 5). Here each class consisting of 400 images of numerals by considering 50% for training and 50% for testing. The performance of an algorithm is tested using 2-fold cross validation when k= 3. The results obtained are encouraging for mixed handwritten and printed Kannada numerals. The Table 3 and Table 4, presents the recognition accuracy of Kannada handwritten and printed Digits separately. The Table 5 shows overall recognition for mixed handwritten and printed Kannada digits.

Table 3: Percentage of Recognition Accuracy for Handwritten Kannada Numerals with KNN and SVM Classifier.

Training samples = 1000, Test samples = 1000 and Number of features = 22				
Handwritten Kannada numeral	**No. of sample Tested**	**No. of sample Trained**	**Percentage of recognition Accuracy with KNN**	**Percentage of recognition Accuracy with SVM**
೦	100	100	97.8	98.2
೧	100	100	98.52	99.01
೨	100	100	97.5	100
೩	100	100	90.25	94.82
೪	100	100	100	100
೫	100	100	98.24	98
೬	100	100	96.58	98.62
೭	100	100	90.45	92.2
೮	100	100	96.82	98
೯	100	100	100.00	100.00
Average Percentage of Recognition accuracy			96.61	97.88

Table 4: Percentage of Recognition Accuracy for Printed Kannada Numerals with KNN and SVM Classifier.

Training samples = 1000, Test samples = 1000 and Number of features = 22				
Printed Kannada numerals	**No. of sample Tested**	**No. of sample Trained**	**Percentage of recognition Accuracy with KNN**	**Percentage of recognition Accuracy with SVM**
0	100	100	98.65	99.8
1	100	100	100	100
2	100	100	97.8	100
3	100	100	100	100
4	100	100	100	100
5	100	100	99.0	100
6	100	100	98.4	99.0
7	100	100	98.8	99,5
8	100	100	99.8	99.8
9	100	100	100	100
Average Percentage of Recognition accuracy			99.245	99.84

Table 5: Percentage of Recognition Accuracy for Mixed Handwritten and Printed Kannada Numerals.

Training samples = 2000, Test samples = 2000 and Number of features = 22				
Mixed Printed and Handwritten Kannada numerals	**No. of sample Tested**	**No. of sample Trained**	**Percentage of recognition Accuracy with KNN**	**Percentage of recognition Accuracy with SVM**
0 ೦	200	200	99.32	99.55
1 ೧	200	200	99.26	99.35
2 ೨	200	200	97.65	100
3 ೩	200	200	95.12	96.03
4 ೪	200	200	100	100
5 ೫	200	200	98.62	98.77
6 ೬	200	200	97.49	100
7 ೭	200	200	94.62	99.018
8 ೮	200	200	98.31	98.54
9 ೯	200	200	100	100

Average Percentage of Recognition accuracy	98.04	99.12

5. Conclusion

An algorithm proposed here for recognition of mixed printed and handwritten Kannada Numerals using Normalized chain code and wavelet transform. The proposed method has shown the encouraging results for recognition of mixed handwritten and printed Kannada digits by using KNN and SVM classifiers. We have obtained recognition accuracy of 98.04% and 99.12% for mixed handwritten and printed Kannada digits with KNN and SVM classifiers respectively. In any recognition process, the important steps are to address the feature extraction method and correct classifier method. The proposed algorithm and classifier tries to meet desired accuracy with few feature vector set. The Novelty of the proposed method is independent of thining.Our future plan is to extend the proposed method for recognition of mixture of printed and hand printed numerals of other Indian languages.

Reffernces

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

[2] N.Sharma,U.Pal,F.Kimura,and S. Pal, "Recognition of Off-LineHandwritten Devanagari Characters Using Quadratic Classifier," ICVGIP 2006, *LNCS* 4338,pp. 805 –816, 2006.

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

[4] B.V.Dhandra, Mallikarjun Hangarge and Gururaj Mukarambi, "Spatial Features for Handwritten Kannada numerals and vowels Character Recognition," *IJCA*, Special Issue on RTIPPR (3):146–151, 2010.

[5] G.Raju , K.Revathy , " wavepackets in the Recognition of Isolated handwritten Malayalam Characters " proceedings of the world Congress on Engineerig Vol 1 , July 2-4 2007

[6] G.G. Rajput, Rajeswari Horakeri "Shape Descriptors Based Handwritten Character Recognition Engine with Application to Kannada Characters", International Conference on Computer & Communication Technology (ICCCT), pp 314-320, 2011.

[7] Kunte Sanjeev R, Sudhaker Samuel (2006), Hu's invariant moments& Zernike moments approach for the recognition of basic symbols in printed Kannada text. Sadhana vol .32, part 5, pp. 521-533. October 2007

[8] G.G.Rajput , Rajeshwari Horakeri, Sidramappa Chandrakant, " Printed and Handwritten Mixed Kannada Numerals Recognition Using SVM" , International journal on Computer Science and Engineering Vol. 02, No.05, pp.1622-1626. ,2010

[9] Gonzales R.C and Woods, R. E. "Digital Image Processing "2nd Ed. Upper Saddle River, N. J, : Prentice-Hall, Inc. (2002)

[10] H. Freeman, Computer Processing of Line Drawings, Computing Surveys, Vol. 6, 57-97

[11] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, 2nd ed., Wiley-New York.

[12] Oivind Trier, Anil Jain, Torfinn Taxt , " A feature extraction method for recognition – a survey " , Pattern Recognition Vol. 29, No 4 . pp 641-662 .1996,

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

[14] B.V. Dhandra , R.G. Benne and Mallikarjun Hangargi, "Handwritten Kannada Numeral Recognition based on structural features " IEEE International conference on Computational Intelligence and Multimedia Application " , ICCIMA-07, pp.157-160 , Dec2007

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

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