Kannada Handwritten Vowels Recognition based on Normalized Chain Code and Wavelet Filters

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

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

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

ABSTRACT
In this paper, an algorithm is designed to recognize the handwritten Kannada vowels based on shape features such as normalized chain codes and wavelet filters. A normalized chain code and wavelet decomposition co-efficients are extracted as a feature vector of size 22 from the normalized binary images of size 40x40 and KNN classifier is applied for classification of vowels. A sample of 1400 Kannada vowels image are used for experiment. The average recognition accuracy of the proposed algorithm is 95.07% for Kannada vowels.

General Terms
Pattern Recognition, Document Image Analysis.

Keywords
OCR, Normalized chain code, wavelet transform, KNN classifier.

1. INTRODUCTION
Development of handwritten Kannada OCR system is one of the challenging area of research due to its diversified applications such as an automatic data entry, mail processing, processing of revenue records and many more. Presently, all communications, businesses transactions are performed through e-technology. Enormous amount of work has been done on OCR for various languages in general and less work on handwritten Kannada documents. Many researchers have developed the character and numeral recognition systems by using template matching, spatial features, Fourier descriptors and shape descriptors, Invariant moments, central moments, Zernike moments and modified invariant moments, structural / statistical, Zoning features and their combinations etc. Different pattern classifiers like neural networks, Hidden markov model, Fuzzy and SVM classifiers. Details of these methods can be found in [3]. Any recognition system of handwritten character should address the problem due to the unconstrained shapes, variation in writing style, intensity of strokes, scale and orientations. To simplify the task, often unconstrained character recognition systems are designed [4]. Extensive work has been carried out for recognition of characters and digits of foreign and Indian languages. In the following section, review of literature is presented for Kannada character recognition.

2. LITERATURE REVIEW
Ashwin et al. [5] have considered the three basic Zones of the Kannada printed character image. Each zone is divided into a number of circular tracks and sectors and feature like the number of ON pixels in each angular region are considered with support vector machine for the classification of characters and achieved an accuracy of 86.11%. U. Pal et al. [6] have used zoning and directional chain code as features vector of size 100 for handwritten Kannada numeral recognition, and obtained reasonably high recognition accuracy. 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. N.Sharma et al. [8] 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. Dhandra et al. [9], have proposed Spatial features viz stroke density, stroke length and the number of strokes as features and k-NN classifier is used to characterize the handwritten Kannada vowels. Their average recognition accuracy reported for handwritten vowel characters is 90.1%. S.Correia et al. [10] have decomposed English numerals image by using Bi-orthogonal wavelets to four sub-images. The feature vector is formed by these wavelet co-efficients and achieved 97.16% recognition accuracy. G.Raju et al. [11] 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%. Dhandra et al. [12], have proposed modified invariant moments features for isolated printed multi font/size kannada vowels and numerals recognition and reported 97.8% accuracy. Rajput et al. [13] 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. Hence, from the above it is clear that the algorithms designed for Kannada characters recognition suffers either from the recognition accuracy or from time and space complexity. Therefore, there is a need to develop an efficient algorithm to recognize the Kannada characters efficiently with minimum number of features and high accuracy as much as possible.

This motivated us to use the normalized chain code and wavelet filter features for all Kannada vowels recognition as the initial step, then an effort will be made for all kannada characters recognition. Hence, special domain (chain code) and frequency domain (Wavelet filter) features are combined to recognize the handwritten Kannada vowels.

Following is the organization of the paper. In Section 3 a details of data Collection and pre-processing is presented.

Section 4 deals with the feature extraction method. The experimental results obtained and comparative analysis are presented in Section 5. Conclusion is given in Section 6.

3. DATA COLLECTION AND PRE-PROCESSING

Standard data base for Kannada character is not available for experimentation. Hence, we have been forced to create our own data base. The data collections are made from the writers belonging to different professions like primary Schools, High Schools and Colleges. These are scanned through a flat bed HP scanner at 300 dpi which usually yields a low noise and good quality document image. The vowels were cropped out 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. The noise as isolated locations and spikes around the end of the vowels are removed using morphological open and close operations. A minimum bounding box is then fitted to the vowels. To bring uniformity among the vowels the cropped vowel image is normalized to 40x40 pixels. A total of 1400 binary images representing Kannada handwritten vowels are obtained. A Sample image of handwritten kannada vowels is shown in Fig 1.

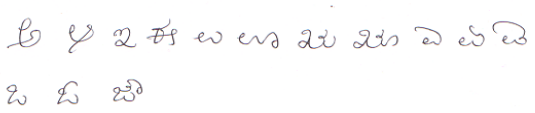

Fig. 1 Sample data set

4. FEATURE EXTRACTION

A well defined feature extraction algorithm makes the classification process very effective and efficient. Two well defined feature extraction methods proposed in this paper are Normalized chain code and wavelet decomposition for recognition of handwritten kannada vowels. A brief description about Normalized chain code and wavelet decomposition are given below.

A. Chain Code

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 [14]. The direction of each segment is coded by using a numbering scheme as shown in Figure 2. According to [15], 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 Algorithm's Step's:

Start:

1. Pick a starting pixel location anywhere on 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 2.
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 right, a code "1" is assigned.
4. The process of locating the next boundary pixel and assigning a code is repeated until we came back to our first location.

End

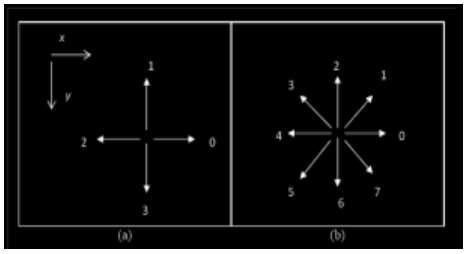

Fig. 2 Direction numbers for (a) 4-directional chain codes, (b) 8-directional chain code

The process of finding the boundary pixel and assigning a code is shown in Figure 3.

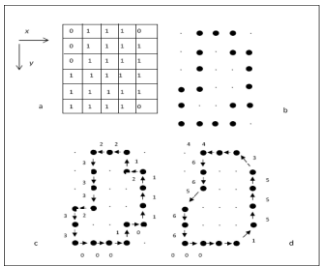

Fig. 3a & b) A 4-connected object and its boundary; c & d) Obtaining the chain code from the object in (a & b) with (c) for 4- connected and (d) for 8-connected

It is observed that the chain code for different vowels has different length code and length of each chain code depends on the size of the handwritten vowels. More ever length of chain code is very high in case of certain handwritten Vowel. We have solved this problem by normalizing the chain code values as explained below. The following chain code is generated for vowels 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 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 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]

Compute the frequency of the codes 0, 1, 2…..7. for vector *V1*

we have the frequency vector V2 as below.

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

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

$$V3 = V2 / |V1| \quad \text{where} \quad |V1| = \sum V2$$

considered the above example, we have

V3= [0.06087 0.08696 0.20000 0.11304 0.21739

0.08696 0.18261 0.05217]

Finally we get the required feature vector of size 08. The algorithm for computing normalized chain codes is given section 4.1

B. 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.

4.1 Algorithm:

Method:

Input:: Isolated and handwritten Kannada preprocessed and normalized vowel image.

Output: Feature library.

Step 1.

Start:

1. Trace the boundary in counterclockwise Direction and generate 8 dimensional chain codes 0 to 7 (Fig.4).
2. Compute the frequency of the codes 0 to 7.
3. Divide frequency of each code by sum of the frequencies.
4. To obtain feature vector of length 08.
5. Repeat the step 1 to step 4 for all sample images representing vowels considered for training.

Step 2.

1. Apply two level forward wavelet packet transform using db4 filter.
2. For each sub band count the number of zero crossings are counted row wise.
3. Obtain the feature vector of length 14.
4. Repeat the step 1 to step 3 for all sample images representing vowels considered for training.

Step 3.

Store the computed normalized chain code from step 1and 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 classifier for classification.

4.2 Algorithm: Recognition of Handwritten Kannada Vowels

Input: Isolated test vowel image.

Output: Recognition of the vowel.

Start:

Step l**:** Extract the features by using Algorithm 4.1.

Step 2**:** Compute the distance between the feature vector 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 vowels.

End.

5. EXPERIMENTAL RESULTS AND DISCUSSION

The proposed algorithm is executed on a database of 1400 isolated handwritten Kannada vowels image from a database created by us (section 3). The database for vowels consists of 1400 samples of the handwritten vowel images, with 100 images representing each class/vowel .For result computation we have used 2-fold cross validation technique. The KNN classifier classifies the test vowels to a class based on K nearest neighbor. The experiments were carried out by varying the values of K i.e. K=1, 3, 5 and found optimal result when K =3. The average recognition rate was obtained for Kannada vowels is 95.07. The classification results of vowels are presented in Table I.

Table I. Average Percentage of Recognition Accuracy for Handwritten Kannada Vowels input with KNN (K=3) Classifiers.

Training samples =700, Test samples =700 , Number of features = 22			
Handwritten Kannada Vowels	No. of sample Trained	No. of Sample Tested	Percentage Of Recognition
	50	50	92
	50	50	100
	50	50	98
	50	50	95
	50	50	96
	50	50	100
	50	50	96
	50	50	98
	50	50	92
	50	50	91
	50	50	90
	50	50	94
	50	50	92
	50	50	97
Average Percentage of Recognition accuracy			95.07

Table II: Experimental Comparative Analysis of Chain code and Wavelet Filter

S.No	Method	Features dimension	Recognition Accuracy in %
1	Chain code	08	77.00
2	Normalized Chain code	08	80.00
3	db filter	14	89.00
4	Chain code and Wavelet filters	22	92.15
5	Normalized Chain code and Wavelet filters	22	95.07

A comparative study of the proposed method with two existing methods proposed by Ashwin and Sastry [5], .Rajput [13] et al. with respect to Character recognition rate, type of features, classifier used is presented in Table III.

Table III. Comparative Analysis

Name	Features computed &Dimension	Classifier used	Character Recognition Rate
Ashwin And Sastry[5]	Zoning features.	Support Vector machine.	86.11%
G.G.Rajput[13]	Chain code and Fourier discripter (608)	SVM classifier	93.92%
Proposed method	Normalised Chain code and Wavelet packet (22)	KNN classifier.	95.08%

6. CONCLUSION

In this paper, a shape based features are proposed for recognition of Handwritten Kannada vowels and to test the performance of an algorithm, 2-fold cross validation is used to recognize handwritten Kannada vowels. The KNN classifier is used to obtain 95.07% recognition accuracy. The experimental analysis of chain code, normalized chain code, db filter, chain code with db filter, normalized chain code with db filter are also compared. The normalized chain code with db filter has given the high recognition accuracy. The future research work is to be extending with other Kannada characters.

7. REFERENCES

[1] R. Plamondon and S. N. Srihari, "On-Line and off-line handwritten recognition: A comprehensive survey," *IEEE Trans on PAMI*, Vol.22, pp 62-84, 2000.

[2] Anil. K. Jain and Torfinn Taxt, "Feature extraction methods for character recognition - A Survey, "Pattern Recognition, vol 29, no.4 .Newyork, pp 641-662, 1996.

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

[4] R M K Sinha, Scott D Conell & Anil K Jain , " Recognition of unconstrained online Devanagari Characters ", proceedings of the 15th International conference on Pattern Recognition , Barcelona, Spain, Sept. 2000, pp. 368-371.

[5] Ashwin T V, Sastry P S 2002 A fonts and size-independen OCR system for printed Kannada documents using support vector machines. Sadhana 27:35-58.

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

[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, October 2007, pp. 521-533.

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

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

[10] S.Correia , J.M Carvallo and R. Sabourin , "On the performance of wavelets for Handwritten Numerals Recognition " . In ICPR 2002 , quebac – Canada, August 2002.

[11] 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.

[12] B.V.Dhandra, Mallikarjun Hangarge and Shashikal Parameshwarappa 2010 , Multi-Font Kannada Vowels Recognition Based on Modified Invariant Moments. Proceedings of RTIPPR- 2010, 119-122.

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

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

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

[16] McAndrew , A. (2004) .Introduction to Digital Image Processing with Matlab. USA: Thomson Course Technology. Pg. 353.

[17] Mr. Anjan Bikash Maity, Mr. Sandip Mandal , Mr. Ranjan Podder, "Edge Detection Using Morphological Method and Corner Detection Using Chain Code Algorithm ", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 4, No 1, July 2011 ISSN (Online): 1694-0814