A Two Stage Approach for Handwritten Kannada Character Recognition

Shashikala Parameshwarppa
Department of Computer Science and Engineering,
Govt. Engg. College, Raichur
Karnataka, India.

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

***Abstract*— This paper presents an efficient Zone based method for recognition of handwritten Kannada characters using two sets of features namely, Second Generation Discrete Curvelet Transform (DCTG2) as the potential features and the density of the object pixels. The proposed algorithm is implemented in two stages. In Kannada character set certain characters are similar in shape, hence such characters are grouped together resulting in 24 classes instead of 48 classes. Images are made noise free by median filter and images are normalized into 64x64 pixels. Curvelet coefficients are used to assign the input character image to one of the groups. In the second stage, object pixel density is used to assign label to the input character image within that identified group. Experiments are performed on handwritten Kannada characters consisting of 9600 images with 200 samples for each character .These features are fed to the KNN classifier for classification of character images. To test the performance of the proposed algorithm two fold cross validation is used. The average recognition accuracy of 92.21% is obtained for Kannada vowels and consonants respectively. The proposed algorithm is independent of thinning.**

Key words: Kannada character Recognition; Curvelets; Standard deviation; KNN classifier.

I. INTRODUCTION

Handwritten character Recognition is one of the important areas in pattern recognition field because it provides solution for document classifications, mail processing, automatic data entry, bank check reading, reading of the customer filled forms and many more. Advancement of e-technology has made the revolution on all fields in general and document automation in particular. This revolution made to develop an OCR system for every languages and scripts for printed and hand printed documents to process automatically. Most of the works related to handwritten character recognition are done in English, Chinese, Japanese and Arabic. The task is more complicated for Indian languages due to complexity in the shape and number of Characters is in similar in shape. The brief summary of the literature is presented below. Several feature extraction techniques are found in the literature for Kannada character recognition. The feature extraction techniques are spatial features, Fourier and shape descriptors, Normalized chain code, Invariant moments, central moments, Zernike moments, modified invariant moments, structural, statistical, Topological, Template Matching, Gabor, Zoning features combinations of these feature etc.

Rakesh Rampalli et al. [1] have proposed fusion of complementary online and offline strategies for recognition of Kannada characters and reported the recognition accuracy of 89.7% with 295 classes. Niranjan et al. [2] have proposed Fisher Linear Discriminant Analysis for unconstrained handwritten Kannada character recognition and reported the recognition accuracy of 57% using angle distance measures. Ragha et al. [3] have used moment based features for recognition of Kagunita (the Kannada compound characters resulting from the consonant and vowel combination).These features are extracted using Gabor wavelets from the dynamically preprocessed original image. Multi-Layer Perception with Back propagation Neural Networks are employed for character classification. Average recognition rate of 86% is reported for vowels and for consonants the average recognition reported is 65%. Aradhya et al. [4] have proposed Fourier transform and principal component analysis technique for handwritten vowels and consonants of Kannada character recognition and achieved the recognition accuracy of 68.89%. For Kannada and English character recognition, Dhandra et al. [5] have used zone based pixel density feature set of size 64 and achieved the 73.33% recognition accuracy for Kannada consonants using SVM classifier. Sanjeev Kunte et al. [6] 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. Dhandra et al. [7] have used Discrete Curvelet Transforms as feature vector for bilingual and trilingual script (Kannada, English and Telgu) identification and reported 94.19%, 95.24% recognition accuracy using Nearest Neighbor classifier. The features used in this algorithm are derived from the Discrete Curvelet Transform (DCVT), introduced by Candes and Donoho in [8]. Dhandra et al. [9], have used second generation discrete curvelet Transform as feature vector for Handwritten Kannada character recognition and reported 90.57% recognition accuracy using KNN classifier. Hence, from the above it is clear that the algorithms designed for Kannada characters recognition are suffers from the recognition accuracy due to Kannada characters are similar shape, the time and space complexity. Hence, there is a need to develop an efficient

algorithm to recognize the Kannada characters effectively with minimum number of features. The observation on characters reveals that the structures of Kannada characters are circles, holes and curvature in nature. This observation made us to use the second generation discrete curvelet transform features to recognize the hand printed isolated Kannada characters .As an initial effort the algorithm is designed for recognition of Kannada vowels and consonants.

Section II of this paper contains the data collection and pre-processing methods. Section III is devoted for feature extraction method and designing of the proposed algorithm for hand printed Kannada Vowels and Kannada consonant character recognition system. The experimental results obtained are presented in Section IV. Comparative analysis is given in Section V and Conclusion is presented in Section VI.

II. DATA COLLECTION AND PREPROCESSING

It is observed that, to validate and verify the results of the proposed algorithm the standard databases for handwritten Kannada character are not available. Hence, the data collection is made and created the own database. Totally 2800 Kannada vowels and 6800 Kannada consonant images are collected from the Varies group of people belonging to 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 consonants 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 isolated character. To bring uniformity among the consonant the cropped consonant image is normalized to 64x64 pixels. A Sample image of the handwritten Kannada Vowels and Kannada consonant is shown in Fig. 1 and Fig. 2.

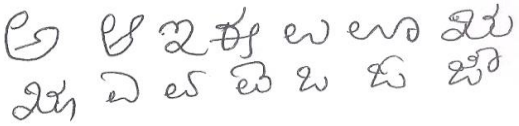


Fig. 1: Handwritten Kannada Vowels

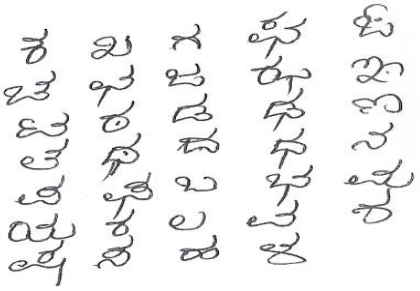


Fig. 2: Handwritten Kannada Consonants

III. FEATURE EXTRACTION

In Kannada character set, few characters are similar in shape and are prone to misclassification. Hence, the similar shape of characters are grouped and labeled as one class together as shown in Fig. 3 and non-similar shape characters are shown in Fig. 4. This resulted in 24 classes as compared to 48 original classes. The details of feature extraction method are given below in two stages.

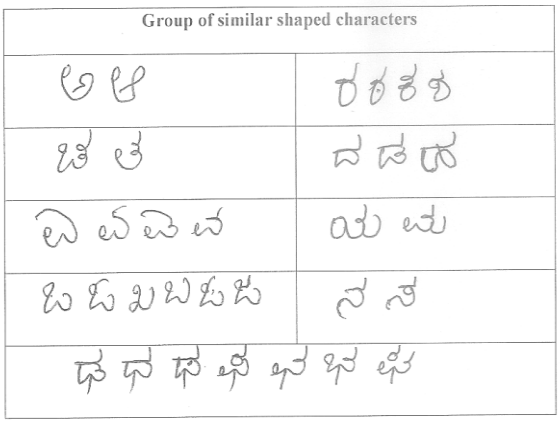


Fig. 3.Group of similar shaped characters.

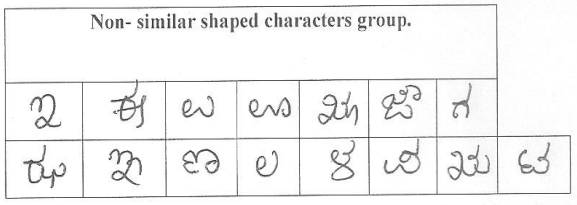


Fig. 4. Non-similar shaped characters.

A. Feature Extraction in the First Stage.

The Kannada handwritten characters have curves and straight lines, so curvelet transform is designed to extract the features, since it allows edges and other singularities along the lines in a more efficient way than other transforms. Hence, in this paper, focus is made on the Discrete Curvelet Transform with the Wrapping Technique. For extracting the features, a wrapping based discrete curvelet transform is used and it can be found in by Candes and Donoho [12]. Curvelet coefficients have different scales and angles. Two parameters are involved in the implementation of curvelet transform: number of scales and number of angles at the coarsest level. Energy of these coefficients is different for different coefficients based on angles and scales. In the proposed method 64X64 image blocks is decomposed into four scales using real-valued curvelets. The number of second coarsest level angles used is 8. After the application of curvelet transform on the input image, one subband at the coarsest and one subband at the finest level of curvelet decompositions are obtained. Different subbands are obtained at each level for the other levels of curvelet decomposition. The number of wedges (subband) is $N_j = 4.2^{\lceil j/2 \rceil}$ at the scale 2^{-j} .When a scale is 1, 2, 3 and 4 then the number of wedges are 4, 8, 16

and 16 respectively. All the coefficients obtained cannot be used in the feature vector as it will increase the size of the feature vector drastically and also the time taken for feature vector formation. Hence, for extracting the potential features and also reducing the size of the feature vector for each sample, the standard deviation is obtained for the first half of the total sub bands at each of the remaining scales except scale 1. Only the first half of the total sub bands are considered, since curvelet angle at θ produces the same coefficients at the angle (θ+π) in the frequency domain. Hence, considering half of the total number of sub bands at each scale reduces the total computation time for the feature vector formation without loss of the information contained in an image. For the finest and the coarsest sub bands the standard deviation calculated is 20 used directly in the feature vector. The feature extraction and recognition process is given in Algorithm- 1 & 2.

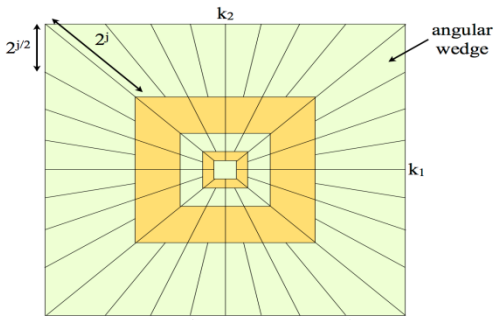


Fig. 5: Rectangular frequency tilling of an image with 5 level curvelets

1. Training Phase

Algorithm-: Feature Extraction Method

Input : Pre processed isolated Handwritten Kannada character image.

Output : Feature library.

Start :

1. Preprocessed image 64X64 pixels
2. Apply Wrapping based discrete Curvelet Transform on the preprocessed image.
3. Different numbers of sub bands are obtained at each level for the other levels of the curvelet decomposition.
4. The scale of 4 and angular orientations 4 are used for 'wedges'. Obtain the curvelets coefficient for each wedge.
5. Compute standard deviations of the curvelet coefficients of the first half of the total subbands (except for scale=1), obtained in step 3 to get feature set of size 20.
6. Repeat the Steps 1 to 5 for all the training images.
7. Computed standard deviations of curvelet coefficients of feature vector size 20, as the features stored in train library in the database

End.

2. Testing Phase.

Algorithm-: Recognition of Handwritten Kannada Character

Input : Isolated test character images.

Output: Recognition of the input Kannada character

Start :

1. Extract the features as obtained in Algorithm-1.
2. Store these feature vectors in test library database.
3. Compute the distance between the feature vectors of the test image stored in the test library and with the feature vector of the trained image stored in the train library.
4. Obtain the minimum distance computed in the step 3. Recognize the character as the label of the train image corresponding to the minimum distances.

End

B. Feature Extraction in second Stage

To discriminate between similar shaped characters in the group (Fig.3), we employed the concept of object pixel density. As discussed in previous paragraph; characters are grouped based on their shape. In each group, we observe that the common part in the image does not contribute to the discriminating features. Hence, we decided to eliminate such common part from the character in the group and compute the features for the remaining part of the character. Examples of such characters are shown in Fig 6. Feature extraction is explained in below.

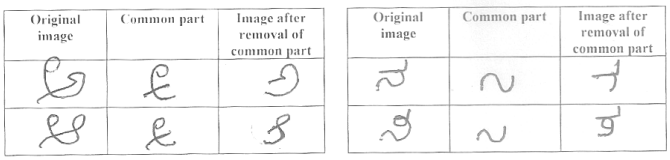
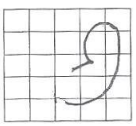


Fig 6. Similar Character's groups after removal of common part



0	0	0	20	7
0	0	0	31	20
0	0	12	8	19
0	0	0	0	4
0	0	11	12	0

Fig. 7. Zoning of a character and object pixel density

The common part of the character is eliminated for a specific group either horizontally or vertically. Traversing the image from top left to bottom right in each zone, the occurrences of object pixel is counted which gives object pixel density for that zone. The feature vector of size 25 is shown in Fig. 7.

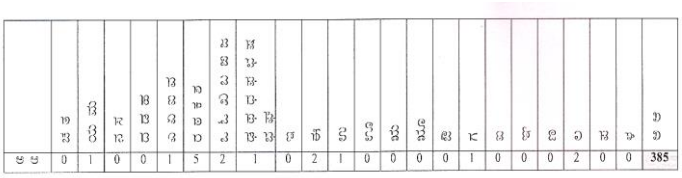
IV EXPERIMENTAL RESULTS

The proposed algorithm is executed on a database of 2800 Kannada vowels and 6800 isolated handwritten Kannada consonants images, with 200 images representing each character. For measuring the performance of an algorithm all preprocessed images are normalized to size 64x64 and experiment is carried out using wrapping based discrete curvelete transform on the preprocessed images. A total of 9600 character images of Kannada characters are classified using KNN classifier. The performance of an algorithm is tested using 2-fold cross validation. The average recognition rate for basic Kannada character is 90.57% from the experiment is presented in our earlier work [9]. The misclassification is mainly due to characters that are similar in shape. Taking into account of similar shaped characters, we then performed experiments in two stages as explained in section 3. Totally 9 groups are formed of 33 similar shaped characters and 15 individual classes of non-group characters, thus resulting in 24 classes instead of 48 classes as shown in Fig 3 and Fig. 4. In the first stage of classification we obtained 93.32% recognition accuracy for 24 classes. Once, the input character was classified to belong to a particular group, the pixel density features of the input character is fed to KNN classifier for character labeling in that group. An example of two stage feature extraction and subsequent classification for the first two vowel characters are shown in Tables 2 to 5. It is clear from the tabulated results that, the method of grouping the characters and then implementing the character recognition in two stages provided acceptable recognition rate of 92.21% compared to that obtained without performing groups (Table 5).

	ಚ	ಐ	ಡ	ಢ	ಷ	ಅ	ಆ	ಇ	ಈ	ಎ	ಐ	ಖ	ಒ	ಓ
ಅ	2	2	1	3	0	176	12	0	0	1	0	2	1	1
ಆ	3	0	2	0	2	8	180	1	2	0	1	0	1	0

Table 2. Confusion matrix of first two vowel characters for 48 classes (not all characters are shown)

Taking into account of similar shaped characters, we then performed experiments in two stages as explained in section 3. Totally 9 groups are formed of 33similar shaped characters and 15 individual classes of non-group characters, thus resulting in 24 classes instead of 48 classes as shown in Fig 3 and Fig. 4. In the first stage of classification we obtained 93.32% recognition accuracy for 24 classes. Once, the input character was classified to belong to a particular group, the pixel density features of the input character is fed to KNN classifier for character labeling in that group. An example of second stage feature extraction and subsequent classification for the first two vowel characters are shown in tables 2 to 5. It is clear from the tabulated results that, the method of grouping the characters and then implementing the character recognition in two stages provided acceptable recognition rate of 92.21% compared to that obtained without performing groups (Table 5).



Stage – 1

	ಅ	ಆ
ಅ	189	05
ಆ	4	190

Stage – 2

Table 3. Confusion matrix of first two vowel characters at (a) Stage-1(obtained from table 4) (b) stage-2

	ಚ	ಐ	ಡ	ಢ	ಷ	ಅ	ಆ	ಇ	ಈ	ಎ	ಖ	ಐ	ಎ	ಓ
ಅ	0	0	2	0	1	189	5	0	1	0	1	0	1	0
ಆ	0	1	1	0	0	4	190	0	0	2	0	1	0	1

Table 4. Confusion matrix of first two vowel characters for 48 classes (consolidated from tables 4, 5)

Class	48 Classes	2 Stage Classification
ಅ	89	94.5
ಆ	90	95
ಇ	92.0	93
ಈ	90.0	90.1
ಉ	92.8	92
ಊ	94.8	96.8
ಎ	90.1	90.1
ಏ	86.4	94.6
ಐ	90.1	90.0
ಒ	94.0	94.6
ಓ	89.0	89.0

Class	48 Classes	2 Stage Classification
ಔ	92.1	92.5
ಕ	86.7	94.5
ಖ	92.0	96.1
ಗ	90.2	94.3
ಘ	88.5	95.2
ಙ	84.5	84.4
ಚ	88.1	92.0
ಛ	90.2	92.0
ಜ	92.5	93.2
ಝ	90.2	91.4
ಞ	88.6	90.3

Class	48 Classes	2 Stage Classification
ಟ	90.4	94.6
ಠ	94.1	89.0
ಡ	90.0	92.7
ಢ	92.0	92.0
ಣ	90.0	95.2
ತ	86.0	92.6
ಥ	94.0	95.0
ದ	90.0	90.0
ಧ	90.0	90.1
ನ	86.0	92.5
ಪ	88.0	94.3

Table 5.Comparison of recognition results of proposed method (Consisting of 24 classes in the first stage) with 48 classes in terms of %

V. COMPARITIVE ANALYSIS.

The Table-6 shows comparative analysis of proposed method with other methods. From the comparative study it is seen that proposed method presents the better recognition accuracy and smaller feature set size as compare to existing other methods.

Authors	Characters Considered	Features Computed & Dimensions	Classifier	Character Recognition Rate
Aradhya et al [4]	Handwritten vowels and Consonants	Fourier transform and PCA	PNN	68.89%
B.V.Dhandra et al [5]	Handwritten Consonants [28 classes]	Zone based Pixel density [64]	SVM	73.33%
Proposed system	Handwritten vowels and Consonants [24 classes]	Step 1:Curvelet Coefficients [20] Step 2:Zone based Pixel density[25]	KNN	92.21%

Table-6: Comparative Analysis of Handwritten Kannada vowels and consonants with other existing Methods.

VI. CONCLUSION

An algorithm proposed here for recognition of handwritten Kannada vowels and Kannada consonants using two sets of feature has exhibited the average percentage of recognition accuracy as 92.21% with KNN classifier with 2-fold cross validation. Two types of feature extraction methods are proposed using curvelet coefficients and pixel density. The proposed method has shown the encouraging results for recognition of Kannada vowels and consonants. This has been demonstrated by performing the experiments on the data set with and without grouping of characters. The aim of the proposed system is to remove the confusions among similar shape characters and thereby increase the recognition rate. The proposed method is to be extended for characters written in other scripts also. The novelty of the proposed method is free from thinning .

REFFERNCES

1. Rampalli R., Ramkrishnan, Angarai G.,(2011) Fusion of Complementary Online and offline Strategies for recognition of Handwritten Kannada Characters Journal of Universal Computer Science, 17(1) . pp 81-93.
2. Niranjan S. K, Vijaya Kumar, Hemantha Kumar(2009) "unconstrained handwritten Kannada character recognition" International Journal of Database Theory and Application, Vol.2, No. 4, pp 290-301
3. Ragha, L. R., Sasikumar, M .: (2011) "Feature Analysis for Handwritten Kannada Kagunita Recognition ". International Journal of Computer Theory and Engineering, IAC-SIT 3(1), pp. 1793-8201
4. Aradhya M., Niranjana S.K., Hemantha kumar G., (2010) "Probabilistic Neural Network based Approach for Handwritten Character Recognition" Special Issue of IJCCT, Vol. 1 Issue 2,3,4 pp. 9-13
5. B.V. Dhandra, Mallikarjun Hangarge and Gururaj Mukarambi (2012). "A Zone Based Character Recognition Engine for Kannada and English Scripts". Elsevier Science Direct, pp. 3292-3299.
6. Kunte Sanjeev R., Sudhaker Samuel (2006). "A simple and efficient optical character recognition system for basic symbols in printed Kannada text". *Sadhana*, Vol. 32, Part 5, pp. 521-533.
7 B.V. Dhandra, Mallikarjun Hangarge, Vijayalaxmi M.B. and Gururaj Mukarambi (2014). "Script Identification Using Discrete Curvelet Transfirms". *IJCA*, Recent Advances in Information Technology. pp.16-20
8 E. Cand`es, L. Demanet, D. Donoho and L. Ying, "Fast Discrete Curvelet Transforms," *Technical Report*, July 2005, pp. 761-799
9 B.V. Dhandra, Shashikala Parameshwarappa (2015). "Handwritten Kannada character Recognition using Curvelet Transform", IJCA, NCDISP2015.pp. 18-24
10 Dinesh Acharya U., N.V. Subba Reddy and Krishnamoorthi (2008). Hierarchical Recognition System for Machine Printed Kannada Characters. IJCSNS *LNCS* International Journal of Computer Science and Network Security Vol. 8 No. 11, pp. 44-53.
11 Nagbhushan P., Pai Radhika M. (1999). Modified region decomposition method and optimal depth decomposition tree in the recognition of non –uniform sized characters – An experimentation with Kannada characters. Pattern Recognition. Letter. 20: pp. 1467-1475
12 E.J. Cand`es and D.L. Donoho, "Curvelets–A Surprisingly Effective Nonadaptive Representation for Objects with Edges," in *Curves and Surfaces*, C. Rabut, A. Cohen, and L.L. Schumaker, Ed., Vanderbilt University Press, Nashville, TN, 2000, pp. 105–120.
13 Ashwin T.V., Sastry P.S. (2002). A Fonts and Size-Independent OCR System for Printed Kannada Documents Using Support Vector Machines. *Sadhana*, 27: pp. 35-58.
14 J.L. Starck, E.J. Cand`es, and D.L. Donoho (2002). "The Curvelet Transform for Image Denoising", *IEEE Trans. Im. Proc.*, Vol. 11, No. 6, pp. 670-684
15 Gonzales R.C. and Woods, R.E. (2002). Digital Image Processing 2nd Ed. Upper Saddle River, N.J.: Prentice- Hall, Inc. pp. 261-269
16 E.J. Candès, L. Demanet, D.L. Donoho, L. Ying. (2003) "Fast Discrete Curvelet Transforms". *Multiscale Model. Simul.*, pp. 861-899.
17 Nagabhushan P., Angadi S.A., Anami B.S. (2003). A Fuzzy Statistical Approach to Kannada Vowel Recognition based on Invariant Moments, Proceedings of NCDAR -2003, PESCE, Mandya, pp. 275-285.
18 U. Pal, B.B. Chaudhuri (2004). Indian Script Character Recognition: A Survey. Pattern Recognition, 37 (2004), pp. 1887–1899.
19 R. Sanjeev Kunte, R.D. Sudhaker Samuel (2007). An OCR System for Printed Kannada Text Using Two-stage Multi-network Classification Approach Employing Wavelet Features, International Conference on Computational Intelligence and Multimedia Applications, pp. 349-355.
20 M.J. Fadili and J.L. Starck (2007). Curvelets and Ridgelets, "Encyclopedia of Complexity and System Science", pp. 1-29
21 Karthik Sheshadri, Pavan Kumar, T. Ambekar, Deeksha Padma Prasad and Dr. Ramakanth P. Kumar (2010). An OCR System for Printed Kannada using K-Means Clustering. IEEE International Conference on Industrial Technology (ICIT), pp. 183-187.
22 Srikanta Murthy, Mamata H.R., Sucharita S., "Multi –font and Multi-size Kannada Character Recognition based on the Curvelet and Standard Deviation", *IJCA*, Vol 35, December 2011. pp.101-104
23 Kumar ,S. Kumar, A ., Kalyan, S.:(2010)"Kannada Character Recognition System using Neural Network". National Journal on Internet Computing. pp. 33-35