Discrete Cosine Transform Coefficients for Kannada Hand-Written Character Recognition

Pateel G P[1]*, Sunil Kumar P[2], Megha N[3]
[1-3] Electronics and Comm. Dept., Sahyadri College of Engineering & Management, Adyar, Mangaluru-575007
*E-mail:pateel.ec@sahyadri.edu.in

ABSTRACT
An offline handwritten Kannada word recognition system using Support Vector Machine (SVM) as classifier is described in this paper. The character recognition system generally involve three major steps viz, preprocessing, feature extraction and classification. In our work in the preprocessing section some of the image processing techniques such as RGB to gray conversion, Binerization, Line segmentation and character segmentation of scanned document are implemented. In the feature extraction section Discrete Cosine Transform coefficients are generated and used to as feature vectors, Later these features (Discrete Cosine Transform) given as inputs to Support Vector Machine (SVM) classifier individually. There by we obtained results.
In order to evaluate the performance of our proposed Optical Character Recognition (OCR) system, 1050 samples of Kannada alphabets written by various people in various styles are made used. Part of this data set is used to train the SVM and remaining part is used to test the performance of SVM. We achieved satisfactory recognition rate of around 86%.

Keywords:
Keywords—Image processing, feature extraction and SVM classification

1. INTRODUCTION

Character recognition is method of recognizing characters from scanned image and converts it into American Standard Code for Information exchange or alternative equivalent machine editable form. This improves the interface between man and machine in numerous applications. In future days, Kannada character recognition system would possibly functions a key issue to form paperless atmosphere by digitizing and process existing paper documents. This method presents an innovative technique to acknowledge written Character. This method will be classified in 2 classes.1) Offline character recognition method 2) Online character recognition method

The offline character recognition method will more split into holistic segmentation approach. In holistic approach word is treated as whole and processed however in segmentation approach every character is separated then processed, in offline recognition method, document is initial created, digitized, hold on in pc then it's processed. Just in case of on--line character identification method, characters square measure processed whereas it's beneath creation. Change of manually written characters is critical for making a few imperative records identified with our history, for example, original copies, into machine editable shape so it can be effectively gotten to and saved. To lessen the exercise in futility associated with composing articles e.g. Kannada daily papers. Valuable under tight restraints processing in banks, all sort of shape preparing frameworks, written by hand post address determination and some more.

Kannada is the official dialect of Karnataka,, More than thirty million individuals talk Kannada as their primary language . Around eleven million individuals utilize Kannada as the second dialect. Kannada has got its own content derived from Brahmi content. Kannada has a base arrangement of 49 Characters. They are ordered into three categories: Swara (vowels), Vyanjana (consonants), and Yogavahakas. There are 13 vowels, 34 consonants and 2 Yogavahakas. The Fig 1.1 represents the Kannada varnamale.

ಅ ಆ ಇ ಈ ಉ ಊ ಋ
ಎ ಏ ಐ ಒ ಓ ಔ ಅಂ ಆಃ

ಕ ಖ ಗ ಘ ಙ
ಚ ಛ ಜ ಝ ಞ
ಟ ಠ ಡ ಢ ಣ
ತ ಥ ದ ಧ ನ
ಪ ಫ ಬ ಭ ಮ

ಯ ರ ಲ ವ ಶ ಷ ಸ ಹ ಳ

2. METHODOLOGY

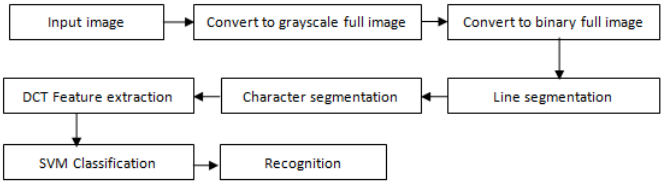


Fig.1: Methodology

2.1 Image Pre-Processing

When an image is fed into the MATLAB handwriting recognition system from the scanner, it is vital to process the image using standard image-processing techniques for easy and appropriate data acquisitions. Noise is the most common portion of the image that has to be discriminated and removed. The following preprocessing techniques as shown in Fig. 1 are used to remove noise and extract individual character handwritten data in MATLAB.

2.1.1 Gray Scale of RGB image

The need for converting to Grayscale is to reduce the processing time for the algorithms; RGB Images are not required to processing.

2.1.2 Removing of Small Object

The gray image contains some small object, so it remove all small object using BWAREAOPEN operation, this operation will remove all small pixel object and it remove small pixel object based on user need and here in our work it removing all small object whose size less than 50 pixel.

2.1.3 Binarization of image

Binarization converts gray-scale image into binary image. During binarization the gray image pixel values with intensity greater than half of the full intensity will be made as '1', which means white and gray image intensity pixel values with intensity less than half of the full intensity will be made as '0' , which means black.

2.1.4 Inversion

Inversion is the process of changing binary image pixel value 1 to 0 which means white color is changed to black and binary image pixel value 0 to 1 which means black color to white. This process is important in extracting a character efficiently from image if it as only one color which is distinct from the background color.

2.2 Segmentation

Segmentation is the process of extraction an individual character from a document this is done in two steps. 1) Line segmentation. 2) Character segmentation. As Kannada is a non cursive script, and individual character in word are isolated. Spacing between the characters can be used for the segmentation. Line segmentation extracts lines from a given image.

Steps to be followed for the line Segmentation is as follows:

1. Scan the image horizontally and identify the non-zero rows between zero rows.
2. Extract the non-zero row from the image that acts as line segment.
3. Repeat the step 1 and 2 for the remaining image until all lines are extracted from the image.

2.3 Feature Extraction

The feature extraction is the process of extracting unique-important properties of an image in the form of feature vector which describes 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. Well known and simple and efficient feature extraction method is Discrete Cosine Transform (DCT) features extraction for handwritten basic Kannada characters recognition system is proposed. A brief description about zoning is given below.

Algorithm for DCT FEATURES

S-1: Read the pre-processed input image

S-2: Compute DCT for the input image (binary image).

S-3: Convert DCT coefficient matrix into 1d zigzag column array.

S-4: Choose the first 50 Discrete-Cosine-Transform (DCT) - coefficients as features.

For a image f (x, y), its second DCT rework is outlined as follows:

$$F(u,v)=\frac{2}{N}C(u)C(v)\sum_{x=0}^{N-1}\sum_{y=0}^{N-1}f(x,y)\cos\left[\frac{(2x+1)u\pi}{2N}\right]\cos\left[\frac{(2y+1)v\pi}{2N}\right] \quad (1)$$

Where

$$\alpha(u),\alpha(v)=\begin{cases}\sqrt{\frac{1}{N}} & \text{For } u,v=0\\ \sqrt{\frac{2}{N}} & \text{Otherwise}\end{cases} \quad (2)$$

Sample Discrete Cosine Transform features extracted for few characters are shown in the following table 1.

Sl. No.				
1	23.12	20.12	18.34	17.5
2	0.855237	1.455498	-2.4183	-1.9366

3	0.114048	2.501746	2.145956	-1.70152
4	0.324902	-1.5406	3.991315	-9.36609
5	0.430636	1.339451	4.524884	0.523709
6	1.480387	2.881773	-3.31648	-3.66928
7	0.739882	0.544732	0.256137	-0.6354
8	6.024383	5.034343	2.005787	-0.21909
9	2.173903	1.552997	1.645417	-1.98539
10	-1.1501	-3.83777	4.097103	-3.2439
11	-2.71565	2.139438	-5.34735	1.144373
12	-0.98202	0.803318	2.527517	0.435731
13	-4.43894	-3.73824	-0.62207	1.969878
14	-1.08616	-2.72938	2.482452	3.357671
15	-1.61976	-4.20701	-1.20904	-2.64659
16	-1.77838	3.878677	4.32767	-0.43862
17	-4.10753	-0.40107	0.937606	-0.9497
18	-3.25963	-3.11966	-0.3425	0.45798
19	1.616914	1.280556	-7.10238	0.583762
20	-0.30078	-0.00612	-1.62683	-0.94998
21	-3.15011	-1.94019	-0.7803	4.289144
22	-3.69199	-4.08389	-1.51703	-0.05086
23	-0.32347	0.839073	-0.20973	-0.32129
24	-4.47463	-6.69927	4.107965	-0.5547
25	0.711191	0.672846	1.057774	2.453092
MEAN	0.091317	0.16367	0.138853	0.193327
VAR	18.52608	13.76344	8.74148	11.09852

The Table-1 represents the Discrete Cosine Transform features extracted for four sample characters which show the difference between mean\variance of one character with the other and the difference is enough to get classification accuracy in case of Discrete Cosine Transform feature extraction method.

2.4 Classification and Recognition

After the feature extraction, the major task is to make decision to classify the character to which class it belongs. There are various classifiers that can apply in recognition. The most important and more effective classifier is Support Vector Machine (SVM).Support vector machine (SVMs) is a supervised learning method used for classification. Where SVM's are a relatively new learning method used for binary classification. The basic idea is to find a hyper-plane which separates the N-dimensional data perfectly into its two classes. SVM commonly used with linear, polynomial, RBF and sigmoid kernels. A multiclass SVM classification has been used in the proposed system with different kernels of 1) linear, 2) polynomial, 3) RBF, 4) sigmoid and it achieves very high recognition accuracy.

The final step is the recognition which is matching the selected class by the SVM with the character and finds the desired character in the Kannada alphabets.

3. Experimental Results and Discussion

The experiments were carried out in Matlab-2015 on a 64-BIT 2.67 GHz INTEL i3 processor, with 2 GB RAM. The dataset consisted of 1050 samples out of which 462 selected samples were used for training and the remaining samples were used for testing. The classification is done using SVM. Fig 3 represents the image sample.

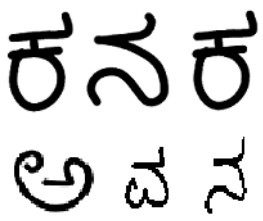


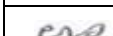
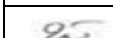
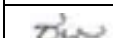
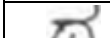
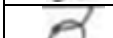
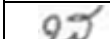
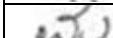
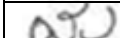
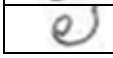
Fig 2 proposed image sample

The following table-2 and 3 gives a summary of the results:

The dataset consisted of 1512 samples out of which 462 selected samples were used for training and 1050 selected samples were used for testing. The classification is done using SVM.

Experiment 1:

Training samples=462 Test samples=1050				
Class	Trained samples	Tested samples	Correct classification	%
	11	25	21	84
	11	25	21	84

	11	25	21	84
	11	25	21	84
	11	25	24	96
	11	25	24	96
	11	25	16	64
	11	25	18	72
	11	25	23	92
	11	25	13	52
	11	25	14	56
	11	25	18	72
	11	25	18	72
	11	25	20	80
	11	25	18	72
	11	25	14	56
	11	25	23	92
	11	25	13	52
	11	25	19	76
	11	25	16	64
	11	25	14	56
	11	25	18	72
	11	25	12	48
	11	25	13	52
	11	25	19	76
	11	25	15	60
	11	25	15	60
	11	25	16	64
	11	25	14	56
	11	25	9	36
	11	25	8	32
	11	25	19	76
	11	25	17	68
	11	25	21	84
	11	25	2	8
	11	25	15	60
	11	25	19	76

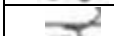
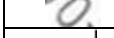
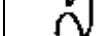
	11	25	14	56
	11	25	14	56
	11	25	4	16
	11	25	20	80
	11	25	16	64
			Average	75.6190 4762

Table-2

Experiment 2:

The dataset consisted of 1512 samples out of which 1050 selected samples were used for training and the remaining samples were used for testing. The classification is done using SVM. The bellow Table-3 represents the percentage of recognition rate of individual characters and the average percentage of recognition is 86%. From the observation of the table 2 and 3 the recognition rate can be improved by using effective feature extraction method and also be improved by using more number of samples for training the SVM.

Training samples=1050 Test samples=462				
Class	Trained samples	Tested samples	Correct classification	%
	25	11	11	100
	25	11	10	90.9090
	25	11	11	100
	25	11	11	100
	25	11	11	100
	25	11	11	100
	25	11	11	100
	25	11	11	100
	25	11	11	100
	25	11	10	90.9090
	25	11	6	54.5454
	25	11	9	81.81818
	25	11	7	63.63636
	25	11	8	72.72727
	25	11	11	100
	25	11	8	72.72727

Training samples=1050 Test samples=462				
Class	Trained samples	Tested samples	Correct classification	%
	25	11	10	90.90909
	25	11	11	100
	25	11	10	90.90909
	25	11	10	90.90909
	25	11	10	90.90909
	25	11	10	90.90909
	25	11	9	81.81818
	25	11	9	81.81818
	25	11	11	100
	25	11	9	81.81818
	25	11	8	72.72727
	25	11	9	81.81818
	25	11	4	36.36363
	25	11	11	100
	25	11	11	100
	25	11	11	100
	25	11	10	90.9090
	25	11	6	54.5454
	25	11	2	18.18181
	25	11	11	100
	25	11	10	90.9090
	25	11	10	90.9090
	25	11	10	90.9090
	25	11	11	100
	25	11	10	90.9090
	25	11	11	100
			Average	86.7965

Table-3

4. CONCLUSION

Recognition of individual character from the handwritten document using image processing techniques, feature extraction methods and finally Support Vector Machine (SVM) as classifier, is implemented in this paper. The recognition rate is around 86%. This work was basically focused on the method that extracts the features efficiently from a single separated character image i.e, Discrete Cosine Transform features and also the method that recognizes the character efficiently i.e, Support Vector Machine (SVM). There by we achieved satisfactory recognition rate for the Kannada hand written words.

The proposed character recognition system of our work can be used to recognize hand written documents of the other languages with suitable modifications. Image de noising and enhancement techniques can be incorporated in the preprocessing section for the degraded image documents.

REFERENCES

1. Rampalli R., Ramkrishnan, Angarai G., “Fusion of Complementary Online and offline Strategies for recognition of Handwritten Kannada Characters “Journal of Universal Computer Science, 17(1) . pp 81-93. 2011
2. Niranjan S. K, Vijaya Kumar, Hemantha Kumar “unconstrained handwritten Kannada character recognition” International Journal of Database Theory and Application, Vol.2, No. 4, pp 290- 301,2009
3. Ragha, L. R., Sasikumar, M.: “Feature Analysis for Handwritten Kannada Kagunita Recognition “. International Journal of Computer Theory and Engineering, IAC-SIT 3(1), pp. 1793-8201 ,2011
4. Aradhya M.,Niranjana S.K.,Hemantha kumar G., “Probabilistic Neural Network based Approach for Handwritten Character Recognition” Special Issue of IJCCT, Vol.1 Issue 2,3,4 pp.9- 13,2010
5. B.V. Dhandra, Mallikarjun Hangarge and Gururaj Mukarambi. “A Zone Based Character Recognition Engine for Kannada and English Scripts”. Elsevier Science Direct, pp. 3292-3299. 2012
6. Kunte Sanjeev R., Sudhaker Samuel. “A simple and efficient optical character recognition system for basic symbols in printed Kannada text”. Sadhana, Vol. 32, Part 5, pp. 521-533. ,2006
7. G.G. Rajput,Rajeswari Horakeri , “ Zone based Handwritten Character Recognition using crack code and SVM ” ,International Conference on advances in Computing, Communication and Informatics(ICACC)-2013
8. S.A.Angadi and Sharanabasavaraj.H.Angadi “structural features for recognition of hand written kannada character based on svm” International Journal of ComputerScience, Engineering and Information Technology (IJCSEIT), Vol. 5,No.2, ,pp.25-32 ,April