Recognition of Similar Handwritten Kannada Consonants.

Akanksha Tonne, Athira AD
B. Tech Students
tonne.akanksha99@gmail.com
athiraasha1126@gmail.com
PES University, Bengaluru

Chandravva Hebbi
Assistant Professor
chandrahebbi@gmail.com
PES University, Bengaluru

Abstract:
Previous efforts on recognition of handwritten character recognition is done on languages like English and Chinese. Less known work is known for recognition of kannada Language. This is due to its complicated script, similarity between characters, and various writing styles. The characters tend to be curvilinear, making it difficult for recognition. This paper makes an attempt to recognize handwritten Kannada characters and analyse how similar characters affect the accuracy of recognition. Our dataset consists of 18000 images of Kannada consonants, each of size 50x50. Each class has approximately 500 images, 36 such classes are considered for our experiment. Popular Machine Learning techniques like SVM, KNN, AdaBoost, Random Forest classifiers with feature extraction methods like SIFT, Zoning are used to do the analysis. Highest accuracy of 72% is obtained for Random Forest classifier with Zoning as the feature extraction method.

Keywords: *Optical Character Recognition, Zoning, SIFT feature, AdaBoost, KNN, Random Forest*

1.Introduction

Handwritten character recognition has been one of the active and challenging research areas in the field of pattern recognition. It can be defined as a process of converting handwritten documents to a machine editable format. In the world of digitisation, character recognition of the handwritten scanned documents is an inviting competition. In the recent past, many new methods and applications of handwritten character recognition are being introduced by researchers. Handwritten Character Recognition is challenging and fascinating area in pattern recognition field. It has various academical and practical applications. It can also be used to process large amount of data thus reducing manual work. Thus, lot of Research has gone into this field for the recognition of English, Chinese and Arabic characters and also some of the Indic Scripts. India is a multi-lingual country with a collection of 22 major languages. Kannada being one such language. Kannadigas, people who are born in Karnataka and speak Kannada, form 72% of the Karnataka's population. It is the official and administrative language of the state and is used in schools, administrative offices, banks and other important places. Its complex script or lipi, is one of the many caveats in learning the language. Kannada language has roughly 43.7 million native speakers due to which, it is the 32nd widely used and spoken language in the world. Kannada is the official language of Karnataka state. It is origined from Kadamba and Chalukya scripts between 5th and 7th century A.D. All the Kannada characters are isolated from each other unlike other Indian languages such as Sanskrit and Hindi that have a line that connects all characters of a word. In the Kannada language different characters can be combined to form compound characters, the number of written symbols is far more than the forty-nine characters in the alphabet.

Modern Kannada has 15 vowels, 34 consonants. The language uses forty-nine phonemic letters, divided into three groups: swaragalu (vowels comprising of 13 letters), vyanjanagalu (consonants comprising of 34 letters), yogavaahakagalu (neither vowel nor consonant comprising of two letters)

Kannada language has been comprehensively and extensively used in many real-world applications, i.e., public sector banks, railways, transport systems, income tax department. Form filling at these places is mandatory and the database is kept in the digital format by scanning the forms and storing them as documents. Absence of recognition system will result in no option for editing of the scanned document and thereby becomes tedious task for correction of the document. Hence, automatic reading system can help save the time.

There is no uniformity in handwritten characters and thus it is challenging to estimate the critical region.

In spite of all these, very little research has been gone in Kannada character recognition. It is due to similarity in many characters which lead to less accuracy. The problem of similar characters can be tackled by extracting additional features from critical region.

As of now, we have worked on consonants and we intend to extend this further for all characters. The Fig 1 shows Kannada Consonants.

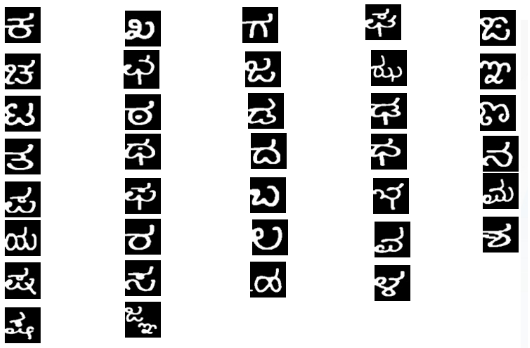


Fig 1. Kannada vyanjanagalu

2. Literature Survey

In [1] Bo Xu et al. have proposed a method for similar handwritten Chinese character recognition by critical region selection. To select critical regions automatically this method used average symmetric uncertainty (ASU) which is a correlation metric used to measure the relevance between a region and the class label. Critical regions are proved to contain more discriminative information and hence can largely benefit the classification for similar characters.

In [2] Swapnil A. Vaidya et al. have described a method of generalized handwritten character recognition using positional feature extraction. Preprocessing methods such as normalization and binarization are used to convert input image into an acceptable form for feature extraction and generalized regression neural network is applied for feature vectors. The proposed recognition scheme provided 82.89% and 85.62% accuracies on Devnagari and kannada character database respectively.

In [3] Reetika Verma et al. have presented an enhanced character recognition method using surf feature and neural network technique. The proposed method has the cabability of strong robustness performance and good distinction between feaure points it have greatly improved in computing speed. The result of the work performed in this paper has the average success rate of 98.7753%. The algorithms have been performed based on noise in input image provide promising results in terms of PSNR and MSE.

In [4] Adwait Dixit, et al. have described a new approach for recognition of handwritten Devanagari characters using wavelet-based feature extraction method. The proposed method used a self-created databse, containing 2000 handwritten characters from different people for training and testing process and obtained an accuracy of 70%.

In [5] Xiwen Qu, Ning Xu et al. have proposed an improved DLA method with few parameters, called adaptive discriminative locality alignment (ADLA) for similar handwritten Chinese character recognition. Compared with DLA, the proposed method has better recognition rate. It inherits all the advantages of DLA and makes the training process become easy due to no computation of parameter optimization.

In [6] Meenu Alex et al., have proposed an approach towards Malayalam Handwritten character recognition using dissimilar classifiers. As an ensemble method, SURF and Curvature features are fed to a neural network and SVM classifier and the output is combined to get final result. The work is divided into two phases first phases uses Malayalam characters and second phase uses Malayalam sentences. An accuracy of 89.9% in phase 1 and 81.1% is observed in phase 2.

In [7] Mahesh Jangid et al., have proposed a way to recognise similar handwritten Devanagari characters by critical region estimation. Starting with an investigation of confusion matrix of similar characters, leading to identification of minor differences in shape, followed by critical region analysis to extract features at the time of classification, is the approach used to solve this problem. Feature extraction methods like Gradient Orientation Detection, and classifiers such as Fisher Linear Discriminant (FLD) and Support Vector Machine (SVM) are used to achieve the result at a 96.58% of accuracy.

In [8] Madhuri Yadav et al., have proposed a way for offline handwritten Hindi character recognition using multiple classifiers namely Quadratic SVM, k-NN, weighted kNN, Ensemble Subspace Discriminant and Bagged Trees. K-fold cross validation technique is used to evaluate the models.

In [9] Sneha Shitole et al., have proposed ways to improve performance of recognition system by using Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). Raw features are extracted by using three different feature extraction methods: chain coding, edge detection, using gradient features and direction feature techniques, which are then reduced by LDA and characters are classified using SVM classifier.

In [10] Prasad K. Sonawane et al., have proposed work on handwritten Devanagari character classification using deep Learning. The network is trained over a dataset of 16870 samples of 22 consonants of Devanagari script. The CNN accuracy and validation are found to be 95.46% and 94.49% respectively.

In [11] Shalini Puri et al., have proposed an efficient Devanagari character classification model using SVM for printed and handwritten mono-lingual Hindi, Sanskrit and Marathi documents. The method first preprocesses the image, segments it through projection profiles, removes shirorekha, extracts features, and then classifies the shirorekha-less characters into pre-defined character categories. The proposed system obtained average classification accuracies of 99.54% and 98.35% for printed and handwritten images, respectively.

From the literature survey we conclude that no previous work includes focus on similar handwritten Kannada Character recognition. In this paper we try to analyse how similar handwritten kannada characters affect the accuracy of model.

3. Proposed System

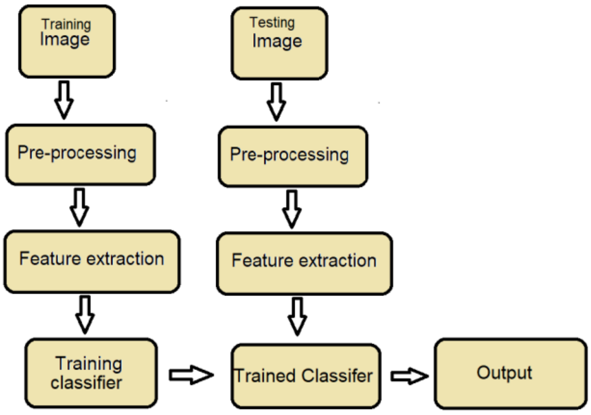


Fig 2. System Architecture

A. Pre-processing

Pre-processing of images is a method which involves removal of noise, reflections and normalizing the intensity of images. Pre-processing is the technique of enhancing data images prior to computational processing. Hence, pre-processing is an important step. Fig 3 shows the scanned image sample before and after pre-processing.

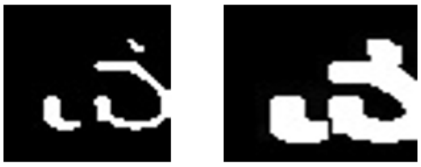


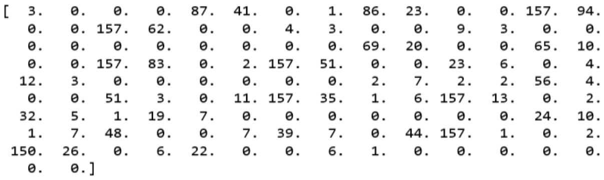
Fig 3: Image sample before and after pre-processing

B. Feature extraction

The **scale-invariant feature transform** (**SIFT**) is a feature detection algorithm in computer vision to detect and describe local features in images [12]. It takes into account scale invariance which is an important property for handwritten character recognition. A 128-dimensional feature vector, Fig 5, is constructed for every key point computed by the SIFT algorithm. Every image has different number of key points, hence different number of descriptors for every image. To make the feature vector size same, K-Means is applied to the set of descriptors for the entire training set. Descriptors are then clustered. The final feature vector, Fig 6, size is equal to the total number of clusters. Every component of the feature vector represents the count of descriptors that got clustered to a bin.

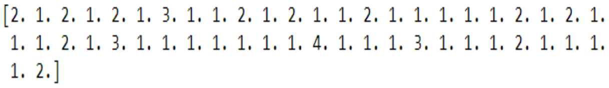
Fig 4. Key points for character ಕ

Example of a discriptor:



```
[  3.    0.    0.    0.   87.   41.    0.    1.   86.   23.    0.    0.  157.   94.
   0.    0.  157.   62.    0.    0.    4.    3.    0.    0.    9.    3.    0.    0.
   0.    0.    0.    0.    0.    0.    0.    0.   69.   20.    0.    0.   65.   10.
   0.    0.  157.   83.    0.    2.  157.   51.    0.    0.   23.    6.    0.    4.
  12.    3.    0.    0.    0.    0.    0.    0.    2.    7.    2.    2.   56.    4.
   0.    0.   51.    3.    0.   11.  157.   35.    1.    6.  157.   13.    0.    2.
  32.    5.    1.   19.    7.    0.    0.    0.    0.    0.    0.    0.   24.   10.
   1.    7.   48.    0.    0.    7.   39.    7.    0.   44.  157.    1.    0.    2.
 150.   26.    0.    6.   22.    0.    0.    6.    1.    0.    0.    0.    0.    0.
   0.    0.]
```

Fig 5. 128-Dimensional Descriptor.



```
[2. 1. 2. 1. 2. 1. 3. 1. 1. 2. 1. 2. 1. 1. 2. 1. 1. 1. 1. 1. 2. 1. 2. 1.
 1. 1. 2. 1. 3. 1. 1. 1. 1. 1. 1. 1. 4. 1. 1. 1. 3. 1. 1. 1. 2. 1. 1. 1.
 1. 2.]
```

Fig 6. Feature vector after clustering the descriptors after K=50.

In **Zone based feature extraction** method the entire image of size 50*50 is divided into zones of size 5*5 and the densities of white pixels in each zone is calculated. .Results are stored in an array and these results (features) are used in the classification methods to recognize the character.

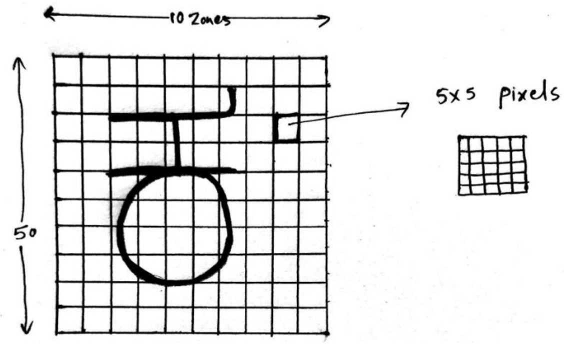


Fig 7. Zone based feature extraction for letter KA

C. Classification

The features described above are tested on multiple classifier namely Support Vector Machine(SVM), K-nearest neighbour(KNN), AdaBoost and Random Forest.

SVM and KNN are supervised learning algorithms. In case of SVM it generates a hyper plane which classifies different classes of training samples.

KNN works based on the concept of distance between the points. It is a instance based learning algorithm where learning stage consists of simply storing the presented training data. When a new query instance is encountered a set of related similar instances are retrieved from the memory and used to classify the new query instance.

AdaBoost and Random Forest are ensembling learning techniques where multiple weak learners are combined to form a strong learner which is better than any of the individual models.

4. Experiment Results

Table-I: shows the number of correctly classified samples for each feature extraction and classifier combination out of 100 testing samples

No	Char	SIFT + SVM	Zoning + KNN	Zoning + AdaBoost	Zoning + Random Forest
1	ಕ	38	77	37	79
2	ಖ	32	87	68	82
3	ಗ	71	81	58	88
4	ಢ	39	66	44	56

5	ಙ	21	76	47	77
6	ಚ	29	75	43	73
7	ಛ	36	73	57	67
8	ಜ	25	72	63	78
9	ಝ	47	96	71	90
10	ಞ	54	93	76	94
11	ಟ	27	81	50	78
12	ಠ	36	62	39	71
13	ಡ	26	54	40	62
14	ಢ	29	67	31	63
15	ಣ	31	80	63	88
16	ತ	29	73	20	73
17	ಥ	17	41	29	46
18	ದ	27	60	14	68
19	ಧ	25	43	42	54
20	ನ	47	74	46	77
21	ಪ	19	58	39	74
22	ಫ	16	44	27	57
23	ಬ	37	82	42	77
24	ಭ	20	71	46	71
25	ಮ	32	78	44	84
26	ಯ	36	83	35	74
27	ರ	47	62	36	73
28	ಲ	83	94	60	94
29	ವ	34	77	51	80
30	ಶ	16	40	39	51
31	ಷ	24	54	37	58
32	ಸ	16	31	31	65
33	ಹ	27	38	18	49
34	ಳ	36	78	45	86
36	ಜ್ಞ	56	98	75	93

SIFT with SVM	30%
Zoning with AdaBoost	44.6%
Zoning with KNN	69%
Zoning with RandomForest	73%

Table II: Average accuracy for each model

Graphical Representation:

Following graphs represent number of misclassified characters for similar characters ಠ (12) and ರ (27) for all the combinations of feature extraction and Classifiers.

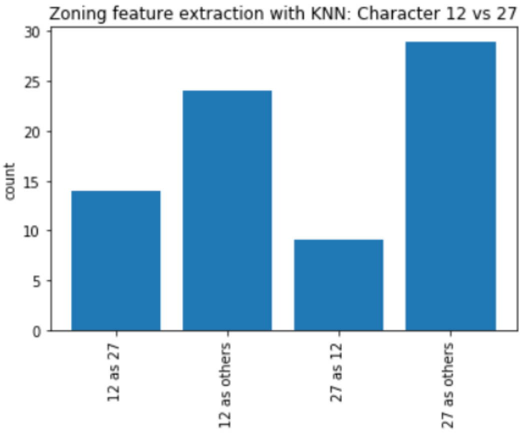


Fig 8. Zoning with KNN.

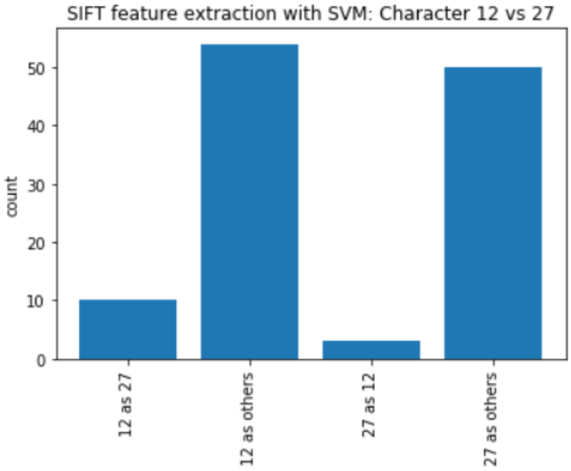


Fig 9. SIFT with SVM.

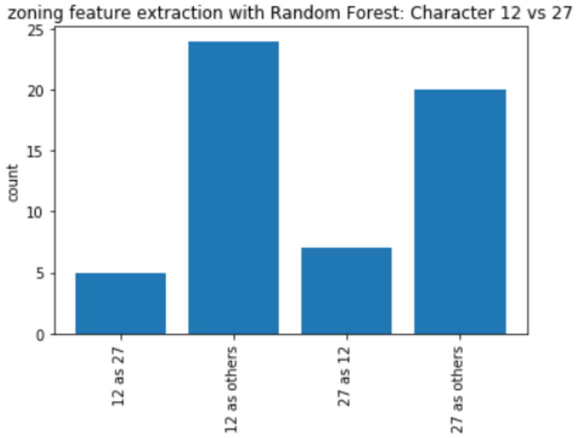


Fig 10. Zoning with Random Forest.

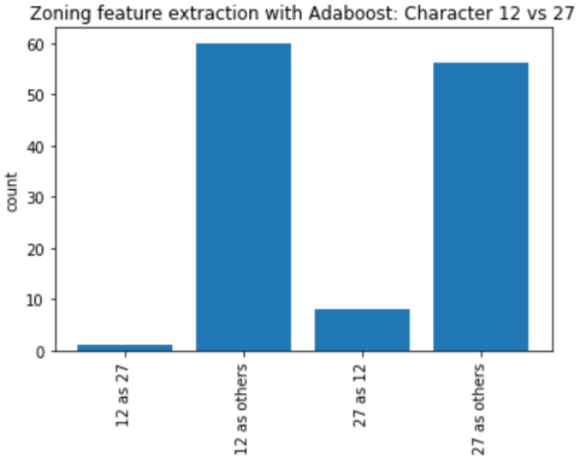


Fig 11. Zoning with Random Forest.

Graphical Interpretation.

The above graphs 8,9,10,11 shows the accuracy analysis performed on similar pair of character ಠ (12) and ರ (27) as an example. They suggest that a high misclassification rate is not only due to presence of similar characters but also due to other reasons like presence of noise in the image.

Conclusion and future enhancements:

After training the models on our dataset, which was split to 80:20 train: test ratio, maximum accuracy of 73% is obtained with Zoning feature extraction method with Random Forest Classifier. Least accuracy was obtained for SIFT feature extraction with SVM classifier. Accuracy for the later method was fluctuating due to random initialization of centroids during K-Means clustering. We plan to improve the accuracy using better clustering methods classification using SIFT feature vectors.

In future our aim is to improve the accuracy of classification by using hierarchical methods, which involve special emphasis on critical regions, which are the region containing more discriminative information between similar characters

References:

[1] Bo Xu, Kaizhu Huang, and Cheng-Lin Liu "Similar Handwritten Chinese Characters Recognition by CriticalRegion Selection Based on Average Symmetric Uncertainty", National Laboratory of Pattern Recognition, 12th International Conference on Frontiers in Handwriting Recognition,2010

[2] Swapnil A. Vaidya and Balaji R. Bombade "A Novel Approach of Handwritten Character Recognition using Positional Feature Extraction", IJCSMC, Vol. 2, Issue. 6, June 2013, pg.179 – 186.

[3] Reetika Verma and Mrs.Rupinder Kaur "Enhanced Character Recognition Using Surf Feature and Neural Network Technique", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (4) , 2014, 5565-5570

[4] Adwait Dixit, Ashwini Navghane, and Yogesh Dandawate "Handwritten Devanagari Character Recognition using Wavelet Based Feature Extraction and Classification Scheme", Annual IEEE India Conference (INDICON), 2014

[5] Xiwen Qu, Ning Xu, Weiqiang Wang and Ke Lu "Similar Handwritten Chinese Character Recognition based on Adaptive Discriminative Locality Alignment", 14th IAPR International Conference on Machine Vision Applications (MVA) May 18-22, 2015. Miraikan, Tokyo, Japan.

[6] Meenu Alexa and Smija Dasb "An Approach towards Malayalam Handwriting Recognition Using Dissimilar Classifiers", Global Colloquium in Recent Advancement and Effectual Researches in Engineering, Science and Technology (RAEREST 2016)

[7] Mahesh Jangid and Dr. Sumit Srivastava, "Similar Handwritten Devanagari Character Recognition by Critical Region Estimation", Intl. Conference on Advances in Computing, Communications and Informatics (ICACCI), Sept. 21-24, 2016.

[8] Madhuri Yadav and Dr. Ravindra Purwar, "Hindi Handwritten Character Recognition using Multiple Classifiers", 7th InternationalConference on Cloud Computing, Data Science & Engineering, 2017.

[9]Sneha Shitole and Savitri Jadhav, "Recognition of Handwritten Devanagari Characters using Linear Discriminant Analysis", Proceedings of the Second International Conference on Inventive Systems and Control (ICISC 2018).

[10] Prasad K. Sonawane and Sushama Shelke, "Handwritten Devanagari Character Classification using Deep Learning", International Conference on Information, Communication, Engineering and Technology (ICICET), 2018.

[11] Shalini Puri and Satya Prakash Singh, “An efficient Devanagari character classification in printed and handwritten documents using SVM”, International Conference on Pervasive Computing Advances and Applications – PerCAA 2019.

[12]David G. Lowe, Distinctive Image Features from Scale-Invariant Keypoints, International Journal of Computer Vision, v.60 n.2, p.91-110, November 2004 [doi>10.1023/B:VISI.0000029664.99615.94]