Machine Learning Approaches for recognition of offline Tulu Handwritten Scripts

C K Savitha 1,  P J Antony 2

1 Department of Computer Science and Engineering, KVGCE, Sullia, DK, VTU, India
2 Department of Computer Science and Engineering, A.J Institute of Engineering & Technology, Mangalore, VTU, India.

1 Email: cksavithaharish@gmail.com, corresponding author
2 Email: antonypjohn@gmail.com

Abstract. This paper presents shallow and deep machine learning techniques for recognizing offline handwritten characters of south Dravidian Tulu scripts. Classification and recognition of characters is carried out using shallow learning techniques like Artificial Neural Networks (ANN), Support Vector Machine (SVM), AdaBoost techniques by extracting zone wise density and gradient features and Deep learning technique called Deep Convolution Neural Network (Deep CNN) classifier. Comparative analysis shows that Deep CNN gives higher efficiency of 98.49% compared with shallow learning techniques for isolated Tulu characters from modern documents and 80.49% for isolated character from Tulu palm leaf manuscripts.

1. Introduction

Character recognition is the method of identifying the input character and changing it into a machine editable format. It offers the machine recognition ability just like people. In optical character recognition (OCR) handwritten texts are converted electronically into system encoded textual content. Handwritten character recognition (HCR) is conducted by two ways, namely online as well as offline. In online processing, the character is processed as soon as it is created. Here, the writer writes with digital device with the use of digital pen. It use stroke, speed, pen-up and pen-down information as parameters. In the case of offline processing the data is processed after its creation. Here, the scanned text document is digitized using a camera or scanner.

The technique by which a computer can recognize every handwritten character is termed as Handwritten Character Recognition system [1]. A good recognition system should handle varieties of writing styles, size, etc. HCR technique can make a digital world by storing whole information safely without any loss. It has major applications in banking, postal field and preserving ancient documents. There are circumstances in a bank, where most of the people while performing transactions may not write the amount legibly in the receipt. Hence, this leads to delay in their money transactions. When such a situation arises, there is a need for good and an efficient HCR system. In the same manner, it can be used in postal fields for recognizing address, amounts, etc. In the case of historical documents, to make textual information of millions of historical documents readily accessible to people, editable form of document is an essential step.

Recently more attention is given in the field of Indian handwritten character recognition. Tulu is one of the five major Dravidian languages in south India. It is an ancient language of Dakshina Kannada and Udupi district of Karnataka State. It has a history of more than 3000 years. Many households in Karnataka have numerous handwritten Tulu manuscripts. As Tulu is not an official language of Karnataka, most of the people are unaware of this language. In order to enhance the readability of Tulu documents, there is a need for machine translation of Tulu scripts into Kannada Script[2].
The basic character set of Tulu consists of 13 vowels and 35 consonants. Tulu vowels with their transcription in Kannada are given in the figure 1. In addition to this basic character set, some dependent and independent vowels, pure consonants and compound characters are also present. All most all characters have curved and loop structure. As we find the similarities in some of the characters, identifying these characters and classifying them to specific class is a tedious job.

The main problem with the HCR system for Indian languages is the lack of dataset and benchmark. The database contains the character set with samples which is created by the programmer. If there is a match in the feature, then the system will recognize that particular character. Thus, character recognition system consists of a group of massive works. The organization of the remaining part of paper is as follows. Related work is discussed in section 2. Section 3 describes the proposed methodologies. Performance comparison with four combinations of techniques has been made in Section 4. Discussion and concluding remarks is explored in Section 5.

2. Related work
Tulu language resembles like Malayalam, various works of HCR for Malayalam language are reviewed in literature [1]. Prime Kumar et al.[3], presented an overview of comparative performance of online Malayalam handwritten character recognition using Hidden Markov Method (HMM) and Support Vector Machine (SVM). When tested on 1279 character samples the system gave maximum accuracy of 97.97% for SVM and 95.24% for HMM. In Nusaibath et al.[4], offline HCR for Malayalam using Gabor filters has been presented. The character was realized based on the extracted feature. The proposed system has recognized both isolated and connected characters with 96.80% recognition accuracy. Malayalam handwritten characters have been successfully recognized using neural network as classifier in [5]. One of the neural networks is trained with 32 gradient features and the other is with 16 density features. For training, back propagation algorithm is used. The final recognition is achieved by combining the results of these individual networks. The system has obtained efficiency of 81.82%. A Multi Protocol Layer Neural Network (MPLNN) based HCR using combined feature extraction is presented in [6]. Neural network is used as hybrid classifier with Genetic Algorithm (GA). A three layered feed forward neural network was used for classification. By combining the four types of features, a recognition rate of 93.23% was obtained. A combined horizontal and vertical projection profile used as feature set and k-Nearest Neighbor (k-NN) and SVM as classifiers in [7] with accuracy of 98.06%. In [8], the curvature feature was extracted using bi-quadratic interpolation method. Support vectors are used for classification of data points. There were ten target classes (0-9). The confusion matrix was built using 90% of training and 10% testing samples which provide 90.5% accuracy for SVM classifier with curvature feature. Signature features are extracted and achieved accuracy of 99% for Mixed National Institute of Standards and Technology (MNIST) data set in [9]. Recurrent neural network based Chinese Character recognition system is proposed in the [10]. Font recognition system on Chinese character has been presented in [11]. Significant performance is achieved for recognition of Offline Chinese Handwriting Characters through the graph based technique in Luo et al [12]. Segmentation lattice creation with generation & scoring of character hypotheses for best path search in the resulting lattice has been described for Multi-Language online Handwriting recognition in [13]. Fusion model of modified quadratic
discriminate function (MQDF) and deep belief network (DBN) has been described to recognize Offline handwritten Chinese character recognition in [14],[15]. Deep learning framework for detection of scene text explained in [16] which gives improved the performance for classification task compared to traditional machine learning approaches. From the literature it has been observed that, for Indian languages, users need to build the database which is a critical task. In case of historical documents, recognition accuracy is poor due to distorted characters with complex backgrounds [17]. There is need to build language dependent machine learning model with suitable classification technique to recognize handwritten characters of regional languages.

3. Methodology
Any character recognition system involves pre-processing, feature extraction, classification and recognition phases. For recognition, a model needs to be developed. The model should be trained first for classification of new test input. For that the character set is divided into training set and testing set. The classifier is built using training samples and the model is validated using testing samples. The proposed system performs comparative study of machine learning models like ANN, SVM, AdaBoost, Deep CNN models for classification and recognition tasks.

The basic function of system is divided into various modules. Each module performs specific function. It is a step by step process where the output from one level is fed as the input to the next level. The main advantage of dividing the functions into module is that the complex problem can be divided into groups of simple problems. Thus, the recognition problem becomes different independent problems. Thus, it enhances the ability to solve the problem. Figure 2 shows various phases involved in system.

![Figure 2. Overview of the system.](image)

3.1. Image Acquisition
Handwritten data is taken as digital input which is acquired using scanning or by taking photograph of Tulu characters. The recognition system needs collection of samples of characters. Samples are collected from different people. When the number of samples is large, higher will be the system accuracy. The acquired samples are grouped into training and testing samples. These two set of samples are then given for pre-processing operation.
3.2. Pre-processing

The main intention of pre-processing is to reduce variations in writing styles of different people. The procedures used in pre-processing steps are noise removal, size normalization and gray scale conversion. Sometimes the input may be colour image. Colour image has different pixel values. So, processing on colour image is a time consuming task. Thus first the entire input colour image is converted into gray scale image. After that the gray scale image is converted into image with only two values i.e., binary conversion. The binary conversion is carried out by Otsu’s global thresholding method. One global threshold value is chosen, for example 160. The pixel value in gray scale image less than 160 is discarded and others are retained. Then only the character region is extracted. The acquired data may contain some unwanted data that may affect further processing of the system. The noise in the acquired data will reduce the system performance. To enhance the quality of the image, filtering techniques such as mask can be applied. To remove the salt and pepper noise, median filter is used. Pre-processing is applied on both trained and testing samples.

3.3. Feature Extraction, Classification and recognition

There is a need to extract set of features that improves the recognition rate with the minimum amount of elements. The extracted feature determines the accuracy of recognizing the characters by the recognition system. The feature set should contain the salient characteristics of the pre-processed image. Most of the Tulu characters are loop and curved structure. Style variations in characters can be captured using statistical distribution points of character image. Zoning is the major technique used to represent character image by statistical features. It may include number of lines in various directions, window height, window width, number of intersection points etc. Both density and directional features are extracted using zone based feature extraction method. Gradient operator is used here to capture direction of intensity changed. Hence, it captures the curved characteristics of the image. In the case of zone based method for obtaining the curvature feature, 8 neighbours of the pixel are considered. For that the image is first converted to 3x3 image by adding zeros to rows and columns if there are no sufficient rows and columns. The features are extracted from these 9 zones as shown in figure 3.

![Figure 3. Tulu character with 3x3 zones.](image)

Contour of each zone is followed and each zone’s 12 directional features are extracted. This feature extraction approach retains the circularity of the normalized image and the normalized data is given to the classifier. The features of characters of both trained samples and the testing set are extracted separately.

Classification is systematic approach to developing model from an input data set. There are two categories as lazy learners and eager learners. Lazy learners delay the process of modelling the training data until classification of test examples. Techniques such as Rote classifiers, k-Nearest Neighbour (k-NN) comes under this category. The lazy learners are also called instance based learning methods take less time in training, but more time in predicting class label and hence it is time consuming task. In the case of eager learning approach, classification model is constructed for given set of training data, before receiving test data to classify. It is committed with single hypothesis that covers the entire instance space; hence it is well suited for character classification. Models such as Artificial Neural Networks (NN), Support Vector Machines (SVM), AdaBoost, and Deep CNN are used to perform recognition task. Here, the extracted features needs to be classified based on their
value. Characters with similar features belong to same class. Thus, the number of class depends on the feature values. Feature values vary based on the character.

Back propagation algorithm is used in ANN to learn weights of output and hidden nodes which is shown in figure 4. It searches for set of weights to minimize mean squared distance between networks class predictions to label actual class of character. The gradient descent method is used here for learning the weights of ANN and to escape from convergence to local minimum, momentum term is added as shown in equation 1.

$$\Delta W_k(i) = -\alpha \frac{\partial E}{\partial w_k} + \mu \Delta W_k(i-1)$$ where $\frac{\partial E}{\partial w_k}$ is the variation of the loss w.r.t. $W_k$  \quad (1)

Where $\alpha$ is the learning rate, and $\mu$ is the momentum term. Training in ANN is time consuming task as the number of hidden nodes is large.

![Figure 4. Multilayer feed forward ANN.](image)

SVM is based on structural risk minimization algorithm and is a statistical method. It works well with high dimensional data. It represents the decision boundary using a subset of training examples known as support vectors. The model is created using SVM as shown in the figure 5.

![Figure 5. SVM linear classifier.](image)

Handwritten character recognition is intrinsically a non linear and high dimensional problem. In order to achieve this non-linear separation, suitable transform function or kernel function must be chosen. Infinite dimensions will not spoil results of maximum margin classifiers as they are well regularized. A number of binary classifiers is combined in One-Against-All multi classifiers, one for each class. Training of each binary classifier is done by taking sample from one of the class as positive and sample from other classes as negative. Then, the binary classifier with greatest output amongst all, corresponding output of multi classifier is activated for the class l. Contrary to SVM; training of neural network tries to minimize the error between the desired output and the output generated by the network.
AdaBoost machine learning model uses output of weighed majority vote of weak classifier to form strong classifier. Quality of training data must be high here. Specifying feature set such as histograms of intensity, direction of gradient are important factors to build strong classifiers. Highly discriminating features are given with higher weightage to classify characters efficiently. Decision stump is created in each iteration on training data. Misclassification is calculated as, Misclassification rate ($M$) = (correct prediction – Training instances) / Training instances. For each training instance weight ($w$) is updated using using equation 2.

$$w = w \cdot \exp (\text{stage} \cdot M)$$

where $	ext{stage}=\ln((1-M) / M)$

In Deep convolution neural networks, the operations are organized into a multilayered with initial layers are used to extract features and fully connected layer at final stage is used to classify characters with soft max activation function as shown in the figure 6.

![Deep CNN Architecture](image)

Figure 6. Deep CNN Architecture.

In all the classification models, each training samples of one character belongs to same class. Similarly, how many different characters are there that much class is defined. The classifiers are first trained according to the train set features and its class labels. The testing is done based on one to all relationship. That is, the test input feature values are checked with all the train feature values. If there is a match then corresponding class value is returned by the model. That is, test set feature values are classified based on the trained model. The recognized character is printed in an editable Kannada text by using equivalent Kannada Unicode for recognized Tulu character. As there is no Unicode for Tulu script, it is mapped into Kannada editable text. The validity of the system is checked by providing input with testing samples. The goal of the system is to predict the correct class of a new data input.

4. Output and performance
Since there is no benchmark data set for Tulu script, 13 vowels and 35 consonant samples for the system are collected from 625 writers. Total 30,000 isolated handwritten Tulu characters are used as dataset. System is trained with 4/5 th of the samples from each character. The recognition ability is checked by giving the characters as input from 1/5 th of the samples. Tulu handwritten samples from palm leaf and modern documents are shown in figure 7, 8.

![Characters from palm leaf](image)

Figure 7. Characters from palm leaf.

![Characters from paper](image)

Figure 8. Characters from paper.
Gradient feature and the density features are extracted from 3×3 zones of each input image. The gradient feature is obtained using sobel operator. The 3×3 sobel operator is applied to each of the 3×3 zones and is considered as a feature set. From each zone, 12 gradient features are extracted. So there are 9×12=108 feature vectors for an image. 108 features of 500 samples of each character is used for training. 108 features of 125 samples of each character is used for testing. Two layer feed forward network with gradient descent learning approach is used in ANN. Back propagation algorithm is applied with learning rate of 0.02, momentum=0.2 and number of epochs=1000 is selected. In the case of SVM based approach, Gaussian kernel with Winner-Takes-All strategy for One-Against-All approach is used. In AdaBoost method, weak classifiers formed by joint probabilities found for which P(feature(i)/Character class is strongly weighed at a place where P(feature(i)/non-character is small. Here learning rate of 0.05 is selected with decision tree as base estimator and n-estimator=5000 is used to terminate the boosting. In case of Deep CNN learning rate of 0.01 is selected with number of epochs=50. For One of the characters recognized by the system is shown in figure 9.

Figure 9. Snapshot of the output for Tulu character.

4.1 Performance Analysis
To evaluate the performance of the classification model, there is a need of counting characters correctly and incorrectly predicted by the model. Confusion matrix table can be used to tabulate the counts. Figure 10 shows the portion of confusion matrix obtained for 125 samples of each character. The summary information achieved using performance metrics such as accuracy, precision, recall, F-score values of the system. Precision value gives the fraction of correctly recognized characters that are relevant. Recall value gives the fraction of relevant elements that are recognized. F-score is expressed in equation 3.

\[ F\text{-score} = \frac{2 \times (\text{precision} \times \text{recall})}{\text{precision} + \text{recall}} \]  

(3)

Figure 10. Confusion matrix plot

To determine which classifier works better on Tulu data set, it is useful to compare the performance of different classifiers. Best classification model is selected based on system with highest...
accuracy, with reducing time complexity when applied to the test set. Precision, recall and F-score values for different classifiers are shown in table 1.

| Dataset                               | Approach | Speed (ms/Character) | Precision | Recall | F-Score |
|---------------------------------------|----------|----------------------|-----------|--------|---------|
| Tulu characters from modern documents | ANN      | 4.01                 | 96%       | 94%    | 94.98%  |
|                                       | SVM      | 2.12                 | 97%       | 95%    | 95.98%  |
|                                       | AdaBoost | 2.34                 | 98%       | 96%    | 96.98%  |
|                                       | Deep CNN | 1.91                 | 99%       | 98%    | 98.49%  |
| Tulu characters from palm leaf manuscripts | ANN      | 4.01                 | 68%       | 72%    | 69.94%  |
|                                       | SVM      | 2.12                 | 69%       | 74%    | 70.46%  |
|                                       | AdaBoost | 2.34                 | 74%       | 72%    | 72.98%  |
|                                       | Deep CNN | 1.91                 | 80%       | 81%    | 80.49%  |

5. Discussion and Conclusion
The proposed handwritten Tulu character recognition system uses zone based gradient operator as feature extraction method for shallow learning techniques such as ANN, SVM and AdaBoost. These techniques fail to achieve refined representation for complex problems. Deep CNN classifier requires less training time as its intermediate layers are used to extract features and final layer to classification. So it outperforms with respect to time and accuracy. The recognition accuracy depends on the quality and number of training samples taken. AdaBoost works well with high quality training data. Its accuracy is reduced in case of damaged characters from Tulu palm leaf manuscripts. Larger the number of samples better the recognition result. The system can recognize characters with various sizes. The efficiency of the system can be improved by applying the novel preprocessing techniques to handle distorted characters. The work can also be extended to word detection.

References
[1] Moni, Bindu S, and G Raju Modified quadratic classifier for handwritten Malayalam character recognition using run length count 2011 ICETECT International Conference on(IEEE) pp 600-604
[2] Poojary Jayakar D Tulu Lipi Terile 2012 A Book on Tulu Script
[3] Prime Kumar K P and Suman Marry Idiculla Online Malayalam Handwritten Character Recognition Using HMM and SVM 2013 ICSIPR(IEEE)
[4] Nusaibath C and Ameera Mol P M Offline Handwritten Malayalam Character Recognition Using Gabor Filters August 2013 IJCTT 4
[5] Anitha Marry M O, Chacko and Dhanya P M Multiple Classifier System for Offline Malayalam Character Recognition 2014 ICICT(Elsevier)
[6] Gauri Katiyar and Shabana Mehfuz MPLNN Based Handwritten Character Recognition Using Combined Feature Extraction 2015 ICCCA(IEEE)
[7] Manoj Kumar Maho, Karamjit Bhatia and R K Sharma A Combined Horizontal and Vertical Projection Feature Extraction 2015 ICACEA IMS Engineering College Ghaziabad India (IEEE)
[8] Pushpalatha Pujari and Babitha Maihi A Comparative Study of Classifiers on Recognition of Offline Odia Numbers 2015 International Conference on EESCO and ICACEA IMS Engineering College Ghaziabad (IEEE)

[9] Lopes, Gustavo Siebra, David Clifte da Silva, Antonio Wendell Oliveira Rodrigues and Pedro Pedrosa Reboucas Filho Recognition of handwritten digits using the signature features and Optimum-Path Forest Classifier 2016 IEEE Latin America Transactions 14(IEEE) 5 pp 2455-2460

[10] Zhang, Xu-Yao, Guo-Sen Xie, Cheng-Lin Liu, and Yoshua Bengio End-to-End Online Writer Identification with Recurrent Neural Network 2017 IEEE Transactions on Human-Machine Systems 47(IEEE) 2 pp 285-292.

[11] Tao, Dapeng, Xu Lin, Lianwen Jin and Xuelong Li. Principal component 2-D long short-term memory for font recognition on single Chinese characters 2016 IEEE transactions on cybernetics 46(IEEE) 3 pp 756-765.

[12] Luo, Yuchen, Rui Xia, and M. Abdulghafour Offline Chinese Handwriting Character Recognition through Feature Extraction 2016 CGiV 13th International Conference on (IEEE) pp 394-398.

[13] Keysers, Daniel, Thomas Deselaers, Henry A. Rowley, Li-Lun Wang and Victor Carbune Multi-language online handwriting recognition 2017 IEEE transactions on pattern analysis and machine intelligence 39(IEEE) 6 pp 1180-1194

[14] Zhong, Zhao, Xu-Yao Zhang, Fei Yin and Cheng-Lin Liu Handwritten Chinese character recognition with spatial transformer and deep residual networks 2016 ICPR 23rd International Conference on (IEEE) pp 3440-3445

[15] Liu, Lu, Weiwei Sun, and Bo Ding Offline handwritten Chinese character recognition based on DBN fusion model 2016 ICIA International Conference on (IEEE)

[16] Tang, Youbao and Xiangqian Wu Scene Text Detection Using Superpixel-Based Stroke Feature Transform and Deep Learning Based Region Classification 2018 IEEE Transactions on Multimedia 20(IEEE) 9 pp 2276-88.

[17] Buric, Jean-Christophe, Mickael Coustaty, Setiawan Hadi, Made Windu Antara Kesiman, Jean-Marc Ogier, Erick Paulus, Kimheng Sok, I.Made Gede Sunarya and Dona Valy ICFHR2016 competition on the analysis of handwritten text in images of balinese palm leaf manuscripts 2016 In Frontiers in Handwriting Recognition ICFHR 15th International Conference on, pp 596-601