A COMPREHENSIVE REVIEW OF ONLINE AND OFFLINE HANDWRITTEN KANNADA CHARACTER RECOGNITION

Abstract

Handwritten Character Recognition (HCR) is an important area of Computer Vision and Image Processing. Algorithms associated with Pattern Matching and Machine learning play a vital role in recognizing the handwritten text or characters. These handwritten characters are stored, preprocessed, required features are extracted, and then the extracted features are trained using Machine learning techniques. Though the few studies describe methods for converting handwritten scripts into system readable format, still there exist a lot of challenges in recognizing handwritten Kannada script. This paper delivers a detailed study of both Online Handwritten Kannada Character Recognition and Offline Handwritten Kannada Character Recognition.

Keywords: Online HCR, Offline HCR, Preprocessing, Feature Extraction, Classifier.

Author

Ajith S
Department of Computer Applications
Nagarajuna College of Management Studies
Chikkaballapura
Karnataka, India.

I. INTRODUCTION

In the era before digitalization, many scripts existed in hardcopy format. Transitioning these handwritten scripts into editable digital forms has been facilitated through image processing techniques. A prominent method used is Handwritten Character Recognition (HCR). This technique allows computers to accept and interpret handwritten input from various sources, such as paper documents, touch-screen devices, photographs, and pen tabs. HCR is divided into two categories: Online HCR and Offline HCR. In the case of Online HCR, Handwritten Characters are recognized while writing through digitizer/pentab or stylus or touchpad. In the case of Offline HCR, Handwritten Characters are recognized from the scanned document. A lot of research has been carried out to recognize handwritten scripts specifically, English scripts. As each person's handwriting is unique it is always challenging to train an algorithm to recognize the characters. Still there exist efficient algorithms to recognize English handwritten characters as the number of characters in English contains 26 uppercase and 26 lower case characters. India has different languages with unique scripts. Especially South Indian regional languages like Kannada, Telugu, Tamil, Malayalam has set of alphabets with different script style. Now the challenge lies in applying character recognition methods to these regional language scripts. When it comes to language like Kannada, it has 49 alphabets consisting of 13 vowels (svara), 2 part-vowel & part-consonant (Yogavaha), 27 structured consonants (Vyanjana), 9 unstructured consonants (Avargeeya Vyanjana). The Kannada script is rich in conjunct consonant clusters, with most consonants having a standard subjoined form and few true ligature clusters. Written Kannada is composed of akshara or kagunita, corresponding to syllables. The letters for consonants combine with diacritics for vowels.

The consonant letter without any diacritic, such as ಕ ka, has the inherent vowel a ಅ This is called dīrgha. A consonant without a vowel is marked with a 'killer' stroke, such as □□ k. This is known as □□□□□□ hrasva [15]. With so many combinations of characters, handwritten character recognition still remains challenging. A sample set of Handwritten Kannada Characters are shown in Figure 1

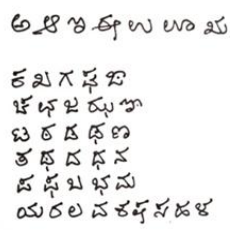

Figure1: Sample Set of Handwritten Kannada Characters.

II. WORKING PRINCIPLE

Generally, the handwritten character recognition technique is divided into five different stages, they are as follows

- Image Acquisition
- Pre-Processing
- Image Segmentation

- Feature Extraction
- Classification

The fundamental stages of HCR are shown in Figure 2

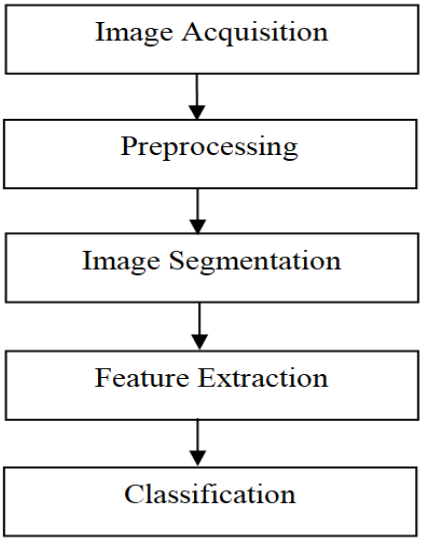

Figure 2: Fundamental Stages of HCR

1. **Image Acquisition:** Image Acquisition is a basic step in HCR that accepts Digital/Digitized images as an input to the system. There are two different ways of image acquisition; the first one is the Offline method which captures an input image for the handwritten character using a digital camera or a scanner. The second method is the Online method that uses Digitizer/Pen Tablet or Stylus, these devices directly accept input to the system using a pen which is digital in nature.

2. **Preprocessing:** The main objective of preprocessing is to improve raw input images so that it reduces the complexity, normalizes the stroke, removes the variation, improves the readability of handwritten characters, and increases the efficiency of an algorithm. The different steps in preprocessing are grayscale conversion, image binarization, noise reduction, size normalization, interpolating missing points, Resampling. Proper preprocessing of raw input images improves the rate of recognition.

3. **Image Segmentation:** Image segmentation intends to divide the image to get the meaningful region so they do not overlap with the unwanted region. The segmentation process is the stage that isolates words and characters from the image. Automating the segmentation improves the recognition rate and efficiency of the HCR system.

4. **Feature Extraction:** The function of feature extraction is to obtain the pattern required for classification. To satisfy the minimum requirement for a good feature extraction there should be minimal intra-class variance and large interclass separation. Feature extraction techniques like Gabor Wavelet, HoG, LBP, Scale Invariant Component Transform (SIFT), Speeded up Robust Features (SURF), ORB, Principal Component Analysis (PCA) can be applied to extract the pattern of handwritten characters. The features which are extracted from handwritten characters can be used further for training the system.

5. **Classification:** Training and Testing are the two stages of Classification. Classification basically analyzes the features of an image and categorizes the data based on its numerical properties. Using these numerical properties, a training class is created by separating the features based on the distinctive description. Some of the Image classification methods that play a vital role in HCR are Convolutional Neural Network, Support Vector Machine (SVM), Kohonen Neural Network, Artificial Neural Network (ANN). Classifier helps to find the best pattern matching for the provided input dataset by comparing the stored pattern with input features.

III. LITERATURE REVIEW

1. **Dutta et al., 2021 [1]**, in this paper SVM multiclass one-to-one classifier technique is carried out to recognize offline Handwritten Kannada letters using Python. The paper also proposes preprocessing stages that play a significant role in reducing the complexity and improving the readability, which involves conversion from RGB image to Grayscale, size reduction, normalization, and feature scaling, and conversion to common-separated values. Support Vector Classifier is employed as an extraction method to classify the object. It was found that using the multiclass-SVM method For a Training data to Test data ratio of 90:10, the efficiency of the algorithm on pre-processed test data was 96.77%. For the Training data to Test data ratio of 80:20, the efficiency of the algorithm on pre-processed test data was 89.1566%. For the Training to Test data ratio of 85:15, the efficiency of the algorithm on pre-processed test data was found to be 95.69%. This technique is implemented on just four letters of the huge Kannada letter sets as of now, and this system can be stretched out to the leftover 49 characters of the letters.

2. **Ramesh et al., 2020 [2]**, have used SVM classifier with CNN and reported an average accuracy of 85% on each input. This technique is used to recognize offline Handwritten Kannada characters. The proposed method employs preprocessing stage in which text is divided into individual words of the entire line. In this paper, the segmentation method based on the Vertical Projection Profile algorithm is used for detecting all the elements in the word image. Though the accuracy is found to be above 80%, the segmentation accuracy is reduced when the spacing between two words is very small.

3. **Sahoo et al., 2020 [3],** used the transfer learning technique to classify offline handwritten characters of the same scriptural family languages like Kannada and Telugu. The significance of this research is to classify a set of two similar character sets. The paper has used Imagenet and COCO dataset. The proposed method reported 84.46% of accuracy with15.54% mean error for the predefined COCO dataset and 88.96% of accuracy with 11.04 % of mean error rate for Imagenet implementation. This can be improved by deploying appropriate pre-processing techniques.

4. **Qisheng Hu, 2020 [4]**, deployed an inception model with data augmentation and transfer learning technique to evaluate Kannada handwritten digit recognition on the Dig-MNIST dataset and obtained a slight gain of 90.27% accuracy over the Kannada-MNIST dataset. Python packages Keras and Tensorflow were used to implement the neural network. The models were trained with a learning rate of 0.001. Better Optimization methods were needed to improve the outcome.

5. **Arun et al., 2019 [5]**, have employed different neural networks for improving the recognition of Kannada characters based on the strokes. The first neural network is CNN-based character recognition which is performed on single character images. The second one is the LSTM-based network that classifies the characters based on stoke sequence. The third one is the combination of CNN-OCR with LSTM-based network, which provides a fully connected layer that classifies the input image. In this paper, Thresholding, Skeletonizing, and Dilation were used as pre-processing techniques before pipelining the data to the network. There are 2174 Kannada data sets that were trained. The proposed method is robust in nature to noise. The combined models of CNN-OCR with LSTM-based networks have gained an accuracy of 97%. However, the accuracy could be improved by using data augmentation.

6. **Sah & Indira, 2018 [6]**, has implemented SVM classifier with RBF kernel to recognize the online Kannada handwritten character. The pen tablet i-ball 5540U was used for the online character acquisition process to capture the data. The paper proposes Smoothening, Resampling, and Normalization as a part of preprocessing techniques. Based on a single stroke, double stroke, and triple stroke, the database consisting of 2940 samples of Kannada characters was constructed. Seven different features were extracted from the data sample, among which normalized coordinates, normalized trajectory, and normalized deviations were selected to test the data ratio based on k-fold cross-validation. The result gave an average recognition rate of 97.14%, 97.55%, and 92.65%, respectively. Only three features were considered among seven features for classification.

7. **Vinod & Niranjan, 2018 [7]**, in this paper, image binarization has played a vital role in recognizing offline Kannada handwritten text. The paper proposes two modules, i.e., Haar wavelet, decomposition, laplacian mask, maximum gradient difference, median filter, and morphological operators, which is implemented as a part of Document Binarization, and Segmentation is implemented using projection profile techniques and skew correction to segment the line. Finally, Connected Component Analysis (CCA) was used to segment the words. The words which were segmented were given as an input to OCR. This paper is implemented using MATLAB R2017a.

8. **Chaithra & Indira, 2017 [8]**, in this paper, i-ball 5540U pen tablet is used to capture samples of Kannada handwritten text to establish the dataset. This paper uses K-Nearest Neighbor (KNN) algorithm with a normalized-based feature extraction method to recognize online handwritten Kannada characters. The performance of the system has gained a recognition rate of 73.47% on a single stroke character set.

9. **Sushma & Veena, 2016 [9]**, has used the Hidden Markov Model-based approach to recognize offline handwritten Kannada words. The proposed system uses maximum entropy sum and entropic correlation techniques to remove the noise from the grayscale image. The output of the word segmentation process and line segmentation process is given as an input to perform feature extraction. The system utilizes Scale-Invariant Component Transform (SIFT), Speeded up Robust Features (SURF), and Feature Extraction and Descriptors (ORB) methods to extract the words. The proposed system has adopted the left-right HMM model as a classifier. RANSAC algorithm is used as a Resampling Technique. OpenCV is used to implement the proposed system. The recognition rate is found to be 32.5% to 75% for each word in a dataset.

10. **Aravinda & Prakash, 2014 [10]**, has developed offline Kannada character recognition which is written by hand. The proposed system recognizes the handwritten characters based on template matching techniques and correlation analysis. In this paper, line segmentation, letter segmentation, and boundary detection is used as part of preprocessing technique. The template matching is used to classify the image. The correlation value of each template image is stored in an array. Matlab (R2009a) software was used to conduct the experiment.

11. **Shwetha & Ramya, 2014 [11]**, this paper reports an online Kannada character recognition system that uses Genius Mouse i608x to collect samples of characters to build a dataset. The smoothing technique is implemented to recognize the characters. The characters were divided into the top stroke, bottom strokes, and middle strokes using the Segmentation technique. The experiments were conducted with KNN and SVM with smoothing techniques and found that the recognition rate is 92.5% and 94.35% respectively.

12. **Keerthi Prasad G et al., 2012 [12],** this paper explains online Handwritten Kannada Character Recognition using principal component analysis approach (PCA) and Dynamic Time Wrapping (DTW). The proposed method is implemented on handheld devices. The rate of recognition based on the PCA approach is found to be 87.5% and the rate of recognition based on the DTW approach is found to be 63.7%. PCA approach has given better results than the DTW approach. However handheld device may not fit due to its computational limitation.

13. **Vishwaas et al., 2012 [13],** the proposed paper explain online handwritten Kannada character recognition using a digitized pen on the touchpad. Recognition mainly involves feature selection using Direction-based Stroke Density Principle (DSD) that determines the density of stroke and records the number of points in every stroke of a character. An unsupervised Kohonen Neural Network (KNN) is used as a classifier. The combined approach of DSD with KNN resulted in delivering an accuracy of 94.4%. The proposed paper is tested on both Kannada handwritten letters and numerical with 20 different handwritten datasets.

14. **Kunte & Samuel, 2000 [14]**, this paper reports an online Kannada character recognition system that uses a digitizer pen tablet for character acquisition. In the proposed paper novel wavelet feature extraction method is used to extract the Kannada character, and a convolutional feedforward multilayer neural network is used as a classifier. The method can be extended to recognize characters of other south Indian languages, specifically, Telugu which is quite similar. The proposed system has reported a recognition rate that is greater than 95%.

IV. CONCLUSION

The paper explains the details of Online and Offline Handwritten Kannada character recognition. The accuracy of recognizing Kannada handwritten characters differs from the nature of the handwritten dataset. In this paper, various methods of handwritten recognition are described. From the study, it is analyzed that, techniques related to classifiers, preprocessing, segmentation, and feature extraction need to be improved to gain a better

recognition rate. The study describes that there is a lot of scope for training the system and enhance the algorithm to improve the rate of recognition.

REFERENCES

[1] K. K. Dutta, S. Arokia Swamy, A. Banerjee, D. R. B, C. R, and D. Vaprani, "Kannada Character Recognition Using Multi-Class SVM Method," pp. 405–409, 2021, doi: 10.1109/confluence51648.2021.9376883.

[2] G. Ramesh, N. Sandeep Kumar, and N. Champa, "Recognition of Kannada Handwritten Words using SVM Classifier with Convolutional Neural Network," 2020 IEEE Reg. 10 Symp. TENSYMP 2020, no. June, pp. 1114–1117, 2020, doi: 10.1109/TENSYMP50017.2020.9231003.

[3] S. Sahoo, B. Prem Kumar, and R. Lakshmi, "Offline handwritten character classification of the same scriptural family languages by using transfer learning techniques," Proc. 3rd Int. Conf. Emerg. Technol. Comput. Eng. Mach. Learn. Internet Things, ICETCE 2020, no. February, pp. 1–4, 2020, doi: 10.1109/ICETCE48199.2020.9091744.

[4] Q. Hu, "Evaluation of deep learning models for kannada handwritten digit recognition," Proc. - 2020 Int. Conf. Comput. Data Sci. CDS 2020, pp. 125–130, 2020, doi: 10.1109/CDS49703.2020.00031.

[5] E. Arun, J. Vinith, C. Pattar, and K. George, "Improving Kannada OCR Using a Stroke-Based Approach," IEEE Reg. 10 Annu. Int. Conf. Proceedings/TENCON, vol. 2019-Octob, pp. 1611–1615, 2019, doi: 10.1109/TENCON.2019.8929606.

[6] R. K. Sah and K. Indira, "Online Kannada Character Recognition Using SVM Classifier," 2017 IEEE Int. Conf. Comput. Intell. Comput. Res. ICCIC 2017, pp. 1–6, 2018, doi: 10.1109/ICCIC.2017.8524435.

[7] H. C. Vinod and S. K. Niranjan, "Binarization and segmentation of kannada handwritten document images," Proc. 2nd Int. Conf. Green Comput. Internet Things, ICGCIoT 2018, pp. 488–493, 2018, doi: 10.1109/ICGCIoT.2018.8753039.

[8] D. Chaithra and K. Indira, "Handwritten online character recognition for single stroke Kannada characters," RTEICT 2017 - 2nd IEEE Int. Conf. Recent Trends Electron. Inf. Commun. Technol. Proc., vol. 2018-Janua, pp. 548–552, 2017, doi: 10.1109/RTEICT.2017.8256657.

[9] A. Sushma and G. S. Veena, "Kannada handwritten word conversion to electronic textual format using HMM model," 2016 Int. Conf. Comput. Syst. Inf. Technol. Sustain. Solut. CSITSS 2016, pp. 330–335, 2016, doi: 10.1109/CSITSS.2016.7779380.

[10] C. V. Aravinda and H. N. Prakash, "Template matching method for Kannada Handwritten recognition based on correlation analysis," Proc. 2014 Int. Conf. Contemp. Comput. Informatics, IC3I 2014, pp. 857–861, 2014, doi: 10.1109/IC3I.2014.7019635.

[11] D. Shwetha and S. Ramya, "Comparison of smoothing techniques and recognition methods for online Kannada character recognition system," 2014 Int. Conf. Adv. Eng. Technol. Res. ICAETR 2014, pp. 0–4, 2014, doi: 10.1109/ICAETR.2014.7012888.

[12] F. K. Keerthi Prasad G, Imran Khan, Naveen R Chanukotimath, "Component Analysis Approach," pp. 675–678, 2012.

[13] M. Vishwaas, M. M. Arjun, and R. Dinesh, "Handwritten Kannada character recognition based on Kohonen Neural Network," Proc. 2012 Int. Conf. Recent Adv. Comput. Softw. Syst. RACSS 2012, pp. 91–97, 2012, doi: 10.1109/RACSS.2012.6212704.

[14] R. S. R. Kunte and R. D. S. Samuel, "On-line character recognition for handwritten kannada characters using wavelet features and neural classifier," IETE J. Res., vol. 46, no. 5, pp. 387–393, 2000, doi: 10.1080/03772063.2000.11416182.

[15] http://brahmi.sourceforge.net/docs/KannadaComputing.html

