

Abstract

The problem of handwritten character recognition (HCR) has been studied extensively over the last few decades by the research community because of their broad applicability. HCR is the process by which a computer converts handwritten text into a machine-readable format. Despite a significant amount of works carried out on HCR of many scripts worldwide, there are some scripts which need to be studied in order to explore script-specific peculiarities and challenges and to validate the existing HCR techniques. Meitei Mayek is of special interest in this regard as the script is newly revived and development of an HCR system will solve real world problems concerning this specific script.

This thesis is a study of HCR of Meitei Mayek. The first problem that the researchers face in the field of HCR of this script is the unavailability of standard datasets. Moreover, development of new datasets opens door for validation of computer vision methods and classification techniques. A publicly available dataset called TUMMHCD of the complete character set of the concerned script has been developed as part of this research work. Performance analysis of four state-of-the-art classifiers with three existing features on the developed dataset is reported in this thesis. Performance analysis is also carried out for five state-of-the-art pre-trained convolutional neural network (CNN) models. A CNN model is proposed for the recognition of TUMMHCD which outperforms the pre-trained models and achieves an accuracy of 95.56%. The proposed CNN also achieves a recognition accuracy of 99.49% on the MNIST dataset.

Using the proposed CNN model, a methodology of multilevel recognition using fusion of features strategy is proposed. It works on the idea that the true class of a misrecognized image is one of the three classes with highest probability values given by the softmax activation function of CNN. In order to identify the final class amongst the three classes, a more discriminative feature set is desired. This feature set is obtained by fusion of deep features and hand-crafted feature descriptors. The performance of the proposed method is validated using different training scenarios, by considering different combinations of the features and also both single-level and multilevel recognition. The multilevel recognition system could enhance the

system performance with a recognition accuracy of 97.06%.

In HCR systems, one of the factors that lead to low recognition accuracy is the presence of confusing characters which are highly shape-similar. This problem is tackled in the present thesis by employing a zone and rule assisted recognition of Meitei Mayek. The proposed methodology identifies the confusing character pairs based on the confusion matrix of the test set of TUMMHCD obtained using the proposed CNN. Script-specific features of Meitei Mayek are then used to distinguish between the characters in each confusing pair. The performance of the proposed method is tested on a Meitei Mayek database of 100 words. It achieves a recognition accuracy of 91.86%.

Incorporation of script-specific features in HCR enhances the system performance. However, limitations pertaining to the adoption of script-specific features cannot be overlooked. The thesis also seeks to develop a methodology to leverage the strengths of CNN and language model (LM) to overcome some of those limitations. The proposed approach considers the decisions given by both the CNN and the LM by combining the class probabilities and conditional probabilities generated by the CNN and the LM respectively. The method achieves a recognition accuracy of 92.29% on the 100 words test dataset.

Keywords: Meitei Mayek, Handwritten character recognition, Dataset creation, Convolutional neural network, Manipuri.