CHAPTER 8
CONCLUSIONS AND FUTURE DIRECTIONS

Handwriting recognition is a vital research area in the field of computer vision, artificial intelligence, and pattern recognition. The text recognition field attained great success in various real-world applications, particularly in the security system, e-government system, and other areas such as cheque recognition, postal address reading for sorting mails, and word spotting on handwritten text sheet. Handwritten word recognition has achieved significant growth since the last two decades, regardless of the languages in which the words are written. To recognize handwritten words, several paradigms have been mentioned in the literature. A lot of work is available in offline handwritten Gurumukhi character recognition but till now no recognized work has been found in offline handwritten Gurumukhi word recognition. Hence, the main objective of this thesis was to propose an offline handwritten Gurumukhi word recognition system which finds its application in postal automation. This goal has been achieved successfully as the system developed recognizes the offline handwritten Gurumukhi words (place names).

The literature related to this work has been incorporated in Chapter 2. Chapter 3 comprises the data collection, digitization, and pre-processing phases for the present work. In Chapter 4, the holistic approach to offline handwritten Gurumukhi word recognition system has been provided. Different features like zoning features, diagonal features and intersection & open-end points features, and also combinations of these features have been employed for the present work. Based on these features, the words have been recognized using three classification techniques, namely, k-NN, RBF-SVM, and Random Forest. To boost the system performance, a majority voting scheme and Adaptive Boosting (AdaBoost) methodology have been employed. Chapter 5 includes the comparative analysis of four feature selection techniques, namely, Consistency Based Analysis (CBA), Correlation Feature Set (CFS), Chi-Squared Attribute (CSA) and Principal Component Analysis (PCA) in order to optimize the boundary extent features extracted from the offline handwritten Gurumukhi words. The classifiers that have been used for this work are Decision Tree and Random Forest. In Chapter 6, Bootstrap Aggregating (Bagging) methodology has

been presented to enhance the performance of this offline handwritten Gurumukhi word recognition system. To extract the features, four feature extraction techniques, namely, zoning features, centroid features, diagonal features, and peak extent features have been considered with their various combinations. For classification, three classification techniques, namely, k-NN, Decision Tree, and Random Forest have been used in this work. Chapter 7 includes the eXtreme Gradient Boosting (XGBoost) technique to encourage the performance of the present offline handwritten Gurumukhi word recognition system based on four feature extraction techniques, namely, zoning features, diagonal features, intersection & open-end points features, and peak extent features. This chapter has been segregated into three sections. Section 8.1 provides a brief contribution to the work. Section 8.2 concludes the present work and section 8.3 focuses on the future scope of the present work.

8.1 BRIEF CONTRIBUTION OF THE WORK

This work has provided the following contributions in the area of offline handwritten Gurumukhi word recognition system.

8.1.1 AdaBoost methodology for offline handwritten Gurumukhi word recognition system

A holistic approach to offline handwritten Gurumukhi word recognition system based on three feature extraction techniques, namely, zoning features, diagonal features, intersection & open-end points features, and three classification techniques, namely, k-Nearest Neighbor (k-NN), Support Vector Machine (SVM), and Random Forest was employed. The system performance was evaluated on a dataset comprising 1,00,000 samples of handwritten Gurumukhi words (place names) based on different combinations of considered features using three evaluation measures such as Accuracy, False Acceptance Rate (FAR), and False Rejection Rate (FRR). Among all the three classification techniques, the Random Forest classifier performed best with a recognition accuracy of 76.15% based on a hybrid of all the three considered features. The classification techniques were combined based on the majority voting scheme that achieved an accuracy of 83.33%, which was observed as superior to the individual classifiers. To improve the recognition performance, AdaBoost

Methodology was applied which achieved maximum accuracy of 88.78% as depicted in Table 4.1 based on a combination of all the three considered features.

8.1.2 Feature selection techniques for offline handwritten Gurumukhi word recognition system

There are numerous feature selection techniques available in the literature to extract the most relevant features and to reduce the dimensionality of the original feature set for pattern recognition. The comparative analysis of four feature selection techniques, namely, Consistency Based Analysis (CBA), Correlation Feature Set (CFS), Chi-Squared Attribute (CSA) and Principal Component Analysis (PCA) for offline handwritten Gurumukhi word recognition has been provided. These feature selection techniques have been employed in this work due to their better recognition results for offline handwritten Gurumukhi character recognition (Kumar *et al.*, 2018). The features extracted from the word images considering the boundary extent feature extraction technique have been reduced by applying feature selection techniques. Two classifiers, namely, Decision Tree and Random Forest have been employed for the classification purpose. The experiments have been evaluated on a public benchmark dataset that comprises 40,000 handwritten Gurumukhi words (place names) using six evaluation measures, namely, Precision, True Positive Rate (TPR), False Positive Rate (FPR), False Rejection Rate (FRR), RMSE (Root Mean Squared Error) and F-Measure. Random Forest classifier attained maximum precision rate (87.32%), maximum TPR (87.42%), minimum FPR (2.28%), minimum RR (9.31%), minimum RMSE (19.50%), and maximum F-measure (87.32%) based on CSA feature selection technique as depicted in Table 5.1. Based on the 5-fold cross-validation technique, the CSA feature selection technique performed best with a recognition accuracy of 87.42% as depicted in Table 5.2.

8.1.3 Bagging Methodology for offline handwritten Gurumukhi word recognition system

In this work, Bagging methodology has been presented to recognize offline handwritten Gurumukhi words in order to get a significant spike in the recognition results. Four feature extraction techniques, namely, zoning features, centroid features, diagonal features and peak extent features (horizontally and vertically), and three

classification techniques, namely, k-NN, Decision Tree, and Random Forest have been employed. Different combinations of features have been considered to evaluate the system performance on the dataset comprising 15,000 samples of handwritten Gurumukhi words (place names). It has been observed that among all the three considered classifiers, the Random Forest classifier performed best with an accuracy of 88.86% based on the combination of zoning, diagonal and vertical peak extent features as depicted in Table 6.1. With the incorporation of Bagging methodology, this rate got improved to 89.92% based on the combination of zoning, centroid, diagonal, and vertical peak extent features as depicted in Table 6.2.

8.1.4 XGBoost technique for offline handwritten Gurumukhi word recognition system

XGBoost technique has been employed for recognition of offline handwritten Gurumukhi words and it was based on four feature extraction techniques, namely, zoning features, diagonal features, intersection & open-end points features, and peak extent features. The system performance has been evaluated on a public benchmark dataset of 40,000 words (place names) based on six evaluation measures such as CPU elapsed time, Accuracy, Precision, Recall, F1-Score, and Area Under Curve (AUC). The dataset has been partitioned into training and testing set using three partitioning strategies (*a*, *b*, and *c*) as depicted in Table 7.1. The proposed system achieved 91.66% (accuracy), 91.39% (precision), 91.66% (recall), 91.14% (F1-score) and 95.66% (AUC) in case of zoning features based on partitioning strategy *a*. The partitioning strategy-wise results based on considered features have been depicted in Tables 7.2-7.5. The comparative analysis of the present approach with the state-of-the-art approaches has been presented in Tables 7.6-7.7 and it has been observed that the present approach attained comparable results.

8.2 CONCLUSIONS

Several techniques such as AdaBoost methodology, Bagging methodology, and XGBoost technique have been employed to boost the performance of the offline handwritten Gurumukhi word recognition system. The maximum accuracies of 88.78%, 89.92%, and 91.66% have been attained based on AdaBoost methodology, Bagging methodology, and XGBoost technique, respectively. Various feature

selection techniques such as CBA, CFS, CSA, and PCA have also been applied to reduce the feature set comprising 128 boundary extent features extracted from the offline handwritten Gurumukhi words. It has been concluded that the CSA feature selection technique performs better as compared to the other three feature selection techniques. A public benchmark dataset in Gurumukhi script has also been created by us which is available for the researchers at the link: https://sites.google.com/view/gurmukhi-benchmark/home/word-level-gurmukhi-dataset that comprises 40,000 offline handwritten words corresponding to 100 different place names collected from 40 different writers.

8.3 FUTURE DIRECTIONS

The work demonstrated in this thesis can be extended in many ways. This section offers some directions following which this work can be extended.

In the present work, the performance of four feature selection techniques, namely, CBA, CFS, CSA, and PCA have been evaluated on the boundary extent feature extraction approach. These feature selection techniques can further be evaluated on different feature extraction approaches as well as their combinations. The present work can further be evaluated based on some latest feature selection techniques in order to improve the existing system performance.

The experimentations have been performed with a single design where each writer wrote each word 10 times. In the future, different designs can be considered in order to evaluate the present offline handwritten Gurumukhi word recognition system. Different partitioning strategies can be considered to partition the dataset into training and testing set so as to achieve optimal accuracy of the system.

In this work, the XGBoost technique has been applied for individual feature extraction techniques. So, one can think of applying various combinations of features to the XGBoost technique for testing the latter's efficiency to the hybrid features. The recognition results can further be enhanced based on different ensemble techniques such as Bayesian model combinations, stacking, etc.

Due to the structural similarity of Gurumukhi script with some other North Indic scripts such as Devanagari, the present work in this thesis can be carried out for the recognition of other Indic scripts.