CHAPTER 5

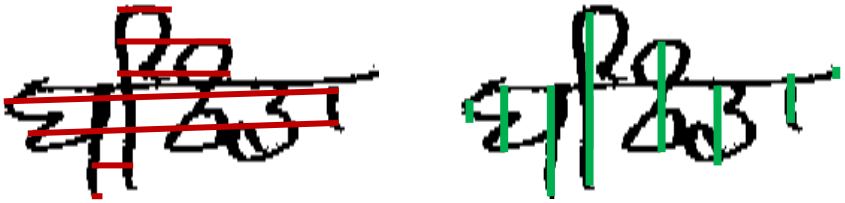
FEATURE SELECTION TECHNIQUES FOR OFFLINE HANDWRITTEN GURUMUKHI WORD RECOGNITION

In certain cases, all the extracted features are not desired to provide training to the classifier. This is because if we employ all the extracted features for the classification purpose, it increases the classification time and also leads to poor system performance due to the existence of irrelevant features in the feature set. Hence, to improve the classification performance, the most significant and relevant features are selected from the original feature set using feature selection techniques. In this direction, four feature selection techniques, namely, Consistency Based Analysis (CBA), Correlation Feature Set (CFS), Chi-Squared Attribute (CSA), and Principal Component Analysis (PCA) have been utilized to optimize the boundary extent features extracted from the word samples. The selected features have been then used to classify the offline handwritten Gurumukhi words based on a holistic approach using two classifiers, namely, Decision Tree and Random Forest. The complete chapter is segregated into 5 sections. Section 5.1 elaborates the boundary extent feature extraction technique and section 5.2 describes the feature selection techniques utilized for the present work. Section 5.3 discusses the experimental results and based on results, the present work has been analyzed in section 5.4. Finally, the summary of the complete chapter is provided in section 5.5.

5.1 BOUNDARY EXTENT FEATURE EXTRACTION TECHNIQUE

The boundary extent feature extraction technique has already been applied for the recognition of offline handwritten Gurumukhi characters (Kumar *et al.*, 2018), which is the motivating factor for the utilization of this technique for the recognition of offline handwritten Gurumukhi words in the present work. This technique extracts the features from the word image based on two ways such as horizontally and vertically as illustrated in Figure 5.1. To extract features horizontally, the word image is segregated into *n* sections in a horizontal way and the boundary extent of the word image is extracted from each section. The boundary extent which is having the largest

length in the particular section is considered as the feature value of that section. The feature value of that section is taken as zero that does not comprise any foreground pixel. This process leads to the *n* element feature set in a horizontal way. Following the same process in a vertical way, *n* element feature set is produced. Hence, a total of 2*n* features is generated comprising both ways i.e. horizontally and vertically. For the present work, the word image has been segregated into 64 sections, and thus, a total of 128 features have been extracted comprising 64 horizontal and 64 vertical boundary extent features.



(a) (b)

Figure 5.1. (a) Horizontal boundary extent feature extraction (b) Vertical boundary extent feature extraction

5.2 FEATURE SELECTION TECHNIQUES

To reduce the feature dimensionality and to select the most significant features from the extracted feature set, various feature selection techniques have been considered such as CBA, CFS, CSA, and PCA which are explored in the following sub-sections:

5.2.1 CBA [Dash and Liu, 2003]

In CBA or Consistency Based Analysis, the inconsistency rate is considered to choose relevant features. An inconsistency gets its origin from instances (0,1,0) and (0,1,1) in which two features contain identical values for the first two instances, but there is a different value in the last instance that represents the class attribute. The smallest feature subset is determined that has the same consistency with respect to the complete feature set.

5.2.2 CFS [Blessie and Karthikeyan, 2012]

In CFS or Correlation Feature Set, the relationship between feature set elements is determined based on the correlation method. Those features get selected which possess less inter-correlation but have higher-correlation with the class.

5.2.3 CSA [Ikram and Cherukuri, 2017]

CSA or Chi-Squared Attribute considers the chi-square test to examine the dependency between the feature and the target (class). Those features are selected which are most dependent on the target; that is, the higher value of the chi-square test specifies that the feature is highly dependent on the target and gets elected for model training.

5.2.4 PCA [Sundaram and Ramakrishnan, 2008]

PCA or Principal Component Analysis is a feature dimensionality approach utilized to find the correlation among a set of variables. It is a mathematical procedure that utilizes a transformation to transform the correlated features into uncorrelated features known as principal components. These principal components are usually less in number as compared to the original variables. In other words, the principal components having maximum variance from one another get selected. Consider P features for handwritten word recognition. The symmetric matrix S representing covariance among these features is computed. Then the eigen vectors Ui (i = 1, 2, 3, … , P) and the complementary eigen values Δi (i = 1, 2, 3, … , P) are computed. From these computed P eigen vectors, the selection of only j eigen vectors is done which correspond to larger eigen values, thus better characteristic features of a word are described. In this way, the features are extracted in PCA based on j eigen vectors.

The feature selection techniques have been employed for the present work due to their following characteristics:

- Minimize the dimensionality of data by incorporating the most relevant features
- Reduce the training time of the model
- Models dependent on feature selection techniques are easy to explain

- Lessen the space requirements
- Foster the system performance
- Improve accuracy

5.3 EXPERIMENTAL RESULTS

The experiments for offline handwritten Gurumukhi word recognition system based on feature selection techniques were evaluated on a public benchmark dataset comprising 40,000 handwritten Gurumukhi words (place names), which is available at the link: https://sites.google.com/view/gurmukhi-benchmark/home/word-level-gurmukhi-dataset (Kaur and Kumar, 2019). To recognize the handwritten words (place names), a holistic approach was used that treats the complete word as an independent unit without considering the segmentation process. The dataset was segregated using an 80:20 partitioning strategy which considered 32,000 (80%) samples for training purpose and the remaining 8,000 (20%) samples to test the present system. To acquire reliable results, a 5-fold cross-validation technique was employed in which the whole data set of each category was divided into 5 equivalent subsets. Out of these 5 subsets, one subset was considered as the testing data and the remaining 4 subsets were considered as the training data. Cross-validation also predicted each sample of the training data. To compare the four feature selection techniques such as CBA, CFS, CSA, and PCA, 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 were used. FRR has already been explained in chapter 4. Rest of the evaluation measures are explored as follows:

- **Precision**

  Precision is interpreted as the proportion of the accurate positive outcomes and the number of predicted positive outcomes as illustrated below:

$$Precision = \frac{TP}{TP+FP} \tag{5.1}$$

  where,

  True Positive (TP): when the true observation is predicted to be true.

  False Positive (FP): when the false observation is predicted to be true.

- **True Positive Rate (TPR)**

  TPR is interpreted as the proportion of positive data samples that are correctly recognized as positive and the total number of positives as illustrated below. It is also known as sensitivity or recall.

$$TPR = \frac{TP}{(FN+TP)} \tag{5.2}$$

  where,

  True Positive (TP): when the true observation is predicted to be true.

  False Negative (FN): when the true observation is predicted to be false.

- **False Positive Rate (FPR)**

  FPR is interpreted as the proportion of negative data samples that are incorrectly recognized as positive and the total number of negatives as illustrated below:

$$FPR = \frac{FP}{(FP+TN)} \tag{5.3}$$

  where,

  False Positive (FP): when the false observation is predicted to be true.

  True Negative (TN): when the false observation is predicted to be false.

- **Root Mean Squared Error (RMSE)**

  RMSE is a widely used measure to detect the variance between the values predicted by a model and the values actually observed. It is also known as Root Mean Squared Deviation (RMSD). The RMSE of an estimator $\widehat{\theta}$ in connection to an estimated parameter θ is interpreted as the square root of the Mean Square Error (MSE) as defined below:

$$RMSE(\hat{\theta}) = \sqrt{MSE}(\hat{\theta}) = \sqrt{E((\hat{\theta} - \theta)^2)} \tag{5.4}$$

- **F-Measure**

  F-Measure, also known as F1-score, is interpreted as the weighted average of the precision and recall as illustrated below. Its value lies in between 0 and 1, 0 being the worst value and 1 being the best value.

$$F - Measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{5.5}$$

For classification based on selected features, two classification techniques, namely, Decision Tree and Random Forest were employed. The performance comparison of four considered feature selection techniques based on evaluation measures is elucidated in Table 5.1.

Table 5.1. Performance comparison of feature selection techniques

Feature Selection Technique	Classifier	Precision	TPR	FPR	FRR	RMSE	F-Measure
CBA	Random Forest	83.36%	83.36%	3.17%	12.47%	22.67%	82.86%
	Decision Tree	51.38%	52.07%	9.41%	37.52%	39.40%	50.99%
CFS	Random Forest	83.75%	83.85%	3.07%	12.08%	22.37%	83.56%
	Decision Tree	64.85%	64.85%	6.83%	27.32%	33.56%	64.45%
CSA	Random Forest	87.32%	87.42%	2.28%	9.31%	19.50%	87.32%
	Decision Tree	76.63%	76.53%	4.46%	18.02%	27.23%	76.33%
PCA	Random Forest	78.41%	77.42%	4.36%	17.23%	26.63%	76.53%
	Decision Tree	71.97%	71.78%	5.45%	21.78%	30.00%	71.78%

Based on the comparative analysis, it has been observed that based on the CSA feature selection technique, the Random Forest classifier attained maximum precision rate (87.32%), maximum TPR (87.42%), minimum FPR (2.28%), minimum FRR (9.31%), minimum RMSE (19.50%) and maximum F-measure (87.32%) and thus, surpassed the other three feature selection techniques as elucidated in Table 5.1. These results are graphically shown in Figure 5.2. In order to test the reliability of the present system, 5-fold cross-validation was performed based on which results are elucidated in Table 5.2. Based on 5-fold cross-validation, CSA feature selection technique attained maximum TPR of 87.42% employing Random Forest classifier.

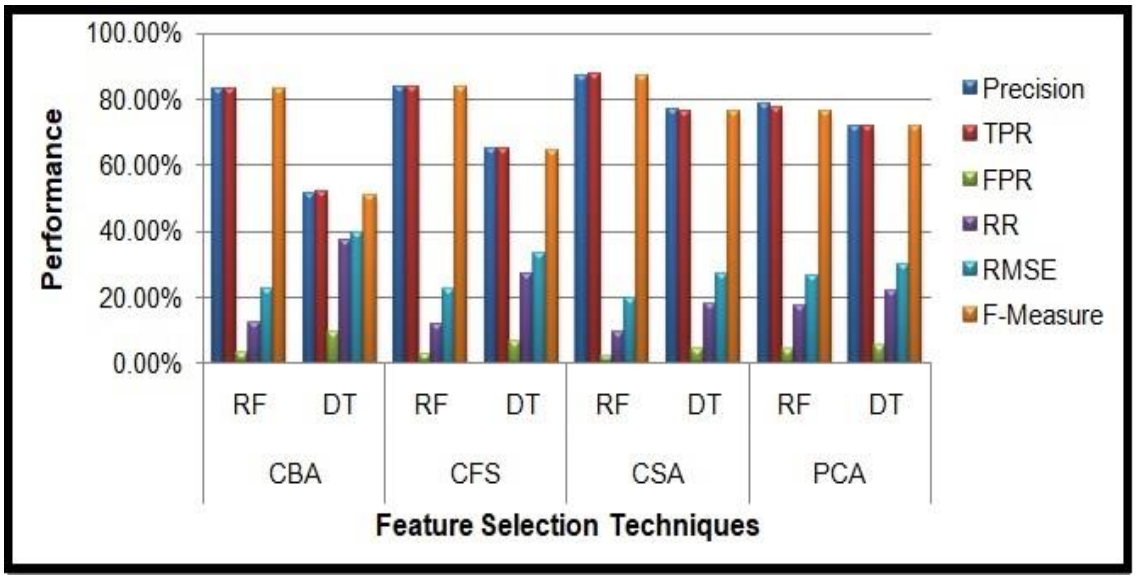


Figure 5.2. Performance comparison of feature selection techniques

Table 5.2. Performance (TPR) comparison of feature selection techniques based on 5-fold cross-validation

Feature Selection Technique	**Classification Technique**	
	Random Forest	**Decision Tree**
CBA	83.36%	52.07%
CFS	83.85%	64.85%
CSA	87.42%	76.53%
PCA	77.42%	71.78%

5.4 ANALYSIS BASED ON EXPERIMENTAL RESULTS

Based on the experimental results, we observed the following key points:

- The experimental approach is a holistic approach that extracts the overall boundary extent features from the complete word without segmenting the word into its primitive components.
- CSA feature selection technique performed best in offline handwritten Gurumukhi character recognition (Kumar *et al.*, 2018), which is also the case in offline handwritten Gurumukhi word recognition.
- Based on each feature selection technique, the maximum results have been attained using the Random Forest classifier due to the fact that it comprises multiple Decision Trees to enhance the classification performance.
- FPR, FRR, and RMSE have been attained less in the case of CSA feature selection technique, hence proving the superiority of CSA as compared to the other three considered feature selection techniques.
- The present work has been evaluated using a public benchmark dataset, hence the experiments conducted can be considered as the baseline references for the future study employing this database.

5.5 CHAPTER SUMMARY

Due to the utilization of all the extracted features, the burden of the classification task can increase in terms of time as well as space. Hence, in this chapter, the offline handwritten Gurumukhi word recognition system based on four feature selection techniques such as CBA, CFS, CSA, and PCA has been presented. These feature selection techniques were employed to reduce the dimensionality of the feature set comprising 128 boundary extent features from the word samples. Based on the selected features, the words (place names) were classified in one of the 100 distinct classes using two classification techniques, namely, Decision Tree and Random Forest. To test the performance of different feature selection techniques, various evaluation measures such as Precision, TPR, FPR, FRR, RMSE, and F-Measure were considered. Based on the considered evaluation measures, the CSA feature selection technique performed best in combination with the Random Forest classifier. The maximum TPR of 87.42% has been achieved by applying the CSA feature selection technique to the Random Forest classifier.