CHAPTER 6

BAGGING METHODOLOGY FOR OFFLINE HANDWRITTEN GURUMUKHI WORD RECOGNITION

Postal automation plays a significant role in the recognition of handwritten addresses on the posting envelopes. In this direction, the offline handwritten word recognition system has been developed to recognize place names handwritten in Gurumukhi script which finds its application in postal automation. For the present work, four feature extraction techniques, namely, zoning features, centroid features, diagonal features and peak extent features (horizontally and vertically) have been considered. To classify the words based on the extracted features, three classifiers, namely, k-NN, Decision Tree, and Random Forest have been employed. Bootstrap Aggregating methodology (Bagging) has also been applied to boost system performance. This chapter is segregated into five sections. Section 6.1 elaborates the processes of different feature extraction techniques used. Section 6.2 explains the working of the Bagging methodology. Section 6.3 demonstrates the experimental results based on Bagging methodology and section 6.4 presents the comparative analysis of the present work with state-of-the-art work. Section 6.5 summarizes the complete chapter.

6.1 FEATURE EXTRACTION TECHNIQUES

In order to extract desirable attributes from Gurumukhi words, four features, namely, zoning features, centroid features, diagonal features and peak extent features (horizontally and vertically) have been considered, which are discussed in the following sub-sections:

6.1.1 Zoning features

In this feature extraction technique, the foreground pixels corresponding to $4^{(L)}$ zones are obtained, where L represents the current level of the word image. For the present work, at first one feature ($4^{(0)}$) was considered from the whole word image. Then the word image was partitioned into 4 zones (4^{1}), which were further partitioned into 4 zones, thus leading to total 4×4=16 zones ($4^{(2)}$). This partitioning continued down to

64 zones ($4^{(3)}$) by splitting each of the 16 zones into 4 zones. It resulted into a total of 1+4+16+64=85 features corresponding to 85 computed zones.

6.1.2 Centroid features

Centroid features are extracted as the center of the word image in the form of (x,y) coordinates. For the present work, initially, the word image was partitioned into 85 number of zones as discussed in the zoning features. Then a total 170 features (85 x-coordinate features and 85 y-coordinate features) were extracted as centroid features from the computed zones of the word image.

6.1.3 Diagonal features

Corresponding to the zones, diagonals features were extracted which were then averaged to attain the single value of each zone. By dividing the word image into 85 zones as illustrated in zoning features, 85 (1+4+16+64=85) diagonal features were extracted for the present work.

6.1.4 Peak extent features

Peak extent features were extracted in two ways i.e. horizontally and vertically. At first, the word image was partitioned into the number of zones as discussed in the zoning features and then 85 (1+4+16+64=85) horizontal peak extent features were extracted from the computed zones. Similarly, 85 vertical peak extent features were extracted from the word image. Thus, in total, 170 peak extent features were extracted for the experimental work.

6.2 BAGGING METHODOLOGY

Bagging or Bootstrap Aggregation, proposed by Breiman (1994), is an ensemble method that combines multiple weak classifiers to get a strong classifier that provides better predictive performance as compared to a single model. In this methodology, the complete training dataset is segregated into multiple subsets which are generated by randomly drawing *N* number of data from the original dataset with replacement, *where N specifies the size of the original training set*. Hence, it develops subsets of the same size as the original dataset. Each of the subsets is utilized to train the individual base classifiers (weak learners) whose predictions are then aggregated based on voting or by averaging to predict the final output. For the present system, a majority voting scheme was employed to form the final prediction as depicted in Figure 6.1.

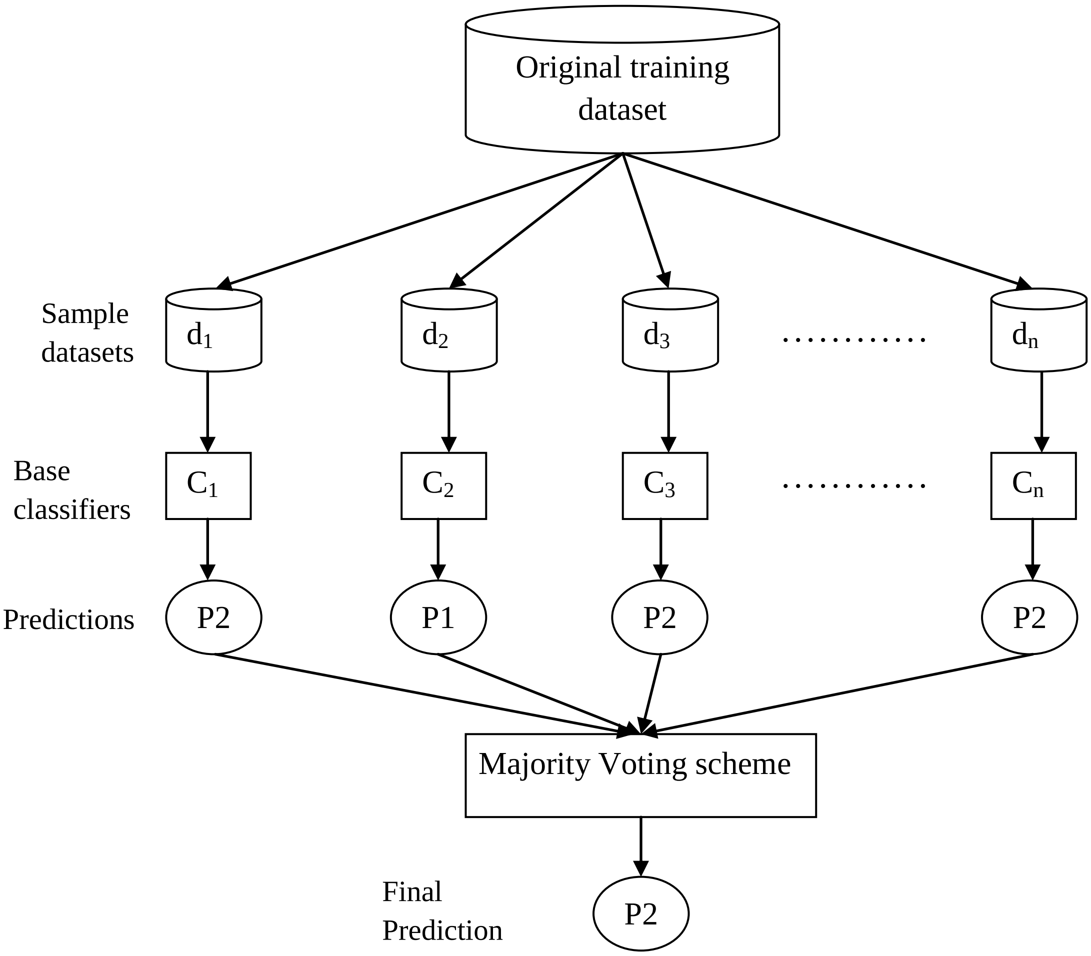

Figure 6.1. Working of Bagging methodology

The Bagging methodology is utilized to boost the system performance due to its following features:

- It performs well for models having high variance such as Decision Trees.
- It aids in minimizing the prediction variance by taking additional data subsets to train the base learners by using bootstrap sampling.
- It reduces model over-fitting by using variance or voting scheme.
- It retains the accuracy for missing data.
- Each model is constructed independently.
- It handles data of higher dimensionality very well.

6.3 EXPERIMENTAL RESULTS

The experiments were conducted on the dataset comprising 15,000 samples of handwritten words in Gurumukhi script that comprises 100 classes of distinct place names. This dataset was gathered from 15 distinct writers, where each writer wrote each word 10 times. The experimental results have been provided based on four feature extraction techniques, namely, zoning features (F1), centroid features (F2), diagonal features (F3) and peak extent features including horizontal peak extent features (F4) and vertical peak extent features (F5) due to popularity and better performance of these features in the literature (Kumar *et al.*, 2013a; Kumar *et al.*, 2013b; Kumar *et al.*, 2014a; Narang *et al.*, 2019). To classify the considered word images, three classification techniques, namely, k-NN, Decision Tree, and Random Forest were considered. The complete dataset was segregated using an 80:20 partitioning strategy in which 12,000 samples (80%) were used to train the present system and the remaining 3,000 samples (20%) were used to test the present system. Table 6.1 demonstrates the recognition results on the basis of the considered features and classifiers. The recognition results based on the Bagging methodology are presented in Table 6.2.

Table 6.1. Recognition results based on the considered features and classifiers without Bagging methodology

Features	Number of feature values	k-NN	Decision Tree	Random Forest
Zoning (F1)	85	72.33%	56.36%	85.17%
Centroid (F2)	170	61.50%	49.80%	82.16%
Diagonal (F3)	85	82.33%	55.56%	86.53%
H_Peak (F4)	85	57.53%	40.13%	70.13%
V_Peak (F5)	85	51.53%	44.56%	72.86%
Zoning+Centroid (F1+F2)	85+170=255	67.50%	55.16%	86.70%
Zoning+Diagonal (F1+F3)	85+85=170	82.36%	55.56%	85.96%
Zoning+H_Peak (F1+F4)	85+85=170	77.83%	56.03%	83.93%
Zoning+V_Peak (F1+F5)	85+85=170	73.23%	56.70%	87.60%
Centroid+Diagonal (F2+F3)	170+85=255	67.60%	54.96%	86.60%
Centroid+H_Peak (F2+F4)	170+85=255	62.73%	47.56%	81.67%

Features	**Number of feature values**	**k-NN**	**Decision Tree**	**Random Forest**
Centroid+V_Peak (F2+F5)	170+85=255	68.83%	50.73%	78.25%
Diagonal+H_Peak (F3+F4)	85+85=170	81.06%	49.70%	82.76%
Diagonal+V_Peak (F3+F5)	85+85=170	74.46%	56.36%	87.50%
H_Peak+V_Peak (F4+F5)	85+85=170	41.86%	46.93%	77.20%
Zoning+Centroid+Diagonal (F1+F2+F3)	85+170+85=340	72.16%	54.83%	86.93%
Zoning+Centroid+H_Peak (F1+F2+F4)	85+170+85=340	71.23%	57.56%	86.66%
Zoning+Centroid+V_Peak (F1+F2+F5)	85+170+85=340	73.73%	57.70%	87.20%
Zoning+Diagonal+H_Peak (F1+F3+F4)	85+85+85=255	84.96%	57.03%	87.23%
Zoning+Diagonal+V_Peak (F1+F3+F5)	85+85+85=255	79.56%	58.63%	88.86%
Zoning+H_Peak+V_Peak (F1+ F4+F5)	85+85+85=255	61.20%	55.30%	84.86%
Centroid+Diagonal+H_Peak (F2+F3+F4)	170+85+85=340	67.83%	48.73%	83.83%
Centroid+Diagonal+V_Peak (F2+F3+F5)	170+85+85=340	74.43%	53.93%	87.30%
Centroid+H_Peak+V_Peak (F2+F4+F5)	170+85+85=340	54.33%	44.76%	77.50%
Diagonal+H_Peak+V_Peak (F3+F4+F5)	85+85+85=255	58.63%	45.63%	82.70%
Zoning+Centroid+Diagonal+ H_Peak (F1+F2+F3+F4)	85+170+85+85=425	74.13%	56.86%	87.93%
Zoning+Centroid+Diagonal+ V_Peak (F1+F2+F3+F5)	85+170+85+85=425	77.63%	58.56%	87.93%
Zoning+Centroid+H_Peak+ V_Peak (F1+F2+F4+F5)	85+170+85+85=425	61.76%	54.70%	84.10%
Zoning+Diagonal+H_Peak+ V_Peak (F1+F3+F4+F5)	85+85+85+85=340	70.03%	54.33%	84.73%

Features	Number of feature values	k-NN	Decision Tree	Random Forest
Centroid+Diagonal+H_Peak+ V_Peak (F2+F3+F4+F5)	170+85+85+85=425	58.83%	45.10%	76.93%
Zoning+Centroid+Diagonal+ H_Peak+V_Peak (F1+F2+F3+F4+F5)	85+170+85+85+85=510	77.66%	55.40%	82.16%

Based on zoning, diagonal, and horizontal peak extent features with k-NN classifier, an accuracy of 84.96% was achieved which remained the same in the case of using the Bagging methodology. Based on the Decision Tree classifier, an accuracy of 58.63% was attained by using zoning, diagonal, and vertical peak extent features. The accuracy attained by the Decision Tree classifier got improved to 81.46% by using the Bagging methodology and considering zoning, centroid, diagonal, and vertical peak extent features. Random Forest classifier attained an accuracy of 88.86% by extracting the zoning, diagonal and vertical peak extent features, which got improved to 89.92% based on Bagging methodology by considering the zoning, centroid, diagonal and vertical peak extent features. Thus, the Bagging methodology proved beneficial to enhance the recognition performance of the system.

Table 6.2. Recognition results based on Bagging methodology

Features	Number of feature values	k-NN	Decision Tree	Random Forest
Zoning (F1)	85	62.23%	78.06%	84.23%
Centroid (F2)	170	67.50%	73.53%	81.03%
Diagonal (F3)	85	82.33%	79.13%	85.16%
H_Peak (F4)	85	57.53%	60.86%	69.50%
V_Peak (F5)	85	51.53%	61.60%	72.36%
Zoning+Centroid (F1+F2)	85+170=255	67.50%	78.26%	86.46%
Zoning+Diagonal (F1+F3)	85+85=170	82.36%	77.40%	85.96%
Zoning+H_Peak (F1+F4)	85+85=170	77.83%	77.36%	82.73%
Zoning+V_Peak (F1+F5)	85+85=170	73.23%	78.30%	86.36%
Centroid+Diagonal (F2+F3)	170+85=255	67.60%	79.20%	87.26%
Centroid+H_Peak (F2+F4)	170+85=255	62.73%	74.41%	81.29%

Features	**Number of feature values**	**k-NN**	**Decision Tree**	**Random Forest**
Centroid+V_Peak (F2+F5)	170+85=255	68.83%	73.96%	82.45%
Diagonal+H_Peak (F3+F4)	85+85=170	81.06%	73.43%	86.41%
Diagonal+V_Peak (F3+F5)	85+85=170	74.46%	79.06%	89.21%
H_Peak+V_Peak (F4+F5)	85+85=170	41.90%	67.36%	77.43%
Zoning+Centroid+Diagonal (F1+F2+F3)	85+170+85=340	72.16%	79.03%	86.24%
Zoning+Centroid+H_Peak (F1+F2+F4)	85+170+85=340	71.23%	79.26%	89.46%
Zoning+Centroid+V_Peak (F1+F2+F5)	85+170+85=340	73.73%	79.96%	85.65%
Zoning+Diagonal+H_Peak (F1+F3+F4)	85+85+85=255	84.96%	81.00%	87.20%
Zoning+Diagonal+V_Peak (F1+F3+F5)	85+85+85=255	79.56%	80.36%	88.27%
Zoning+H_Peak+V_Peak (F1+F4+F5)	85+85+85=255	61.23%	77.33%	84.13%
Centroid+Diagonal+H_Peak (F2+F3+F4)	170+85+85=340	67.83%	73.73%	79.18%
Centroid+Diagonal+V_Peak (F2+F3+F5)	170+85+85=340	74.43%	79.06%	86.12%
Centroid+H_Peak+V_Peak (F2+F4+F5)	170+85+85=340	54.33%	68.53%	74.25%
Diagonal+H_Peak+V_Peak (F3+F4+F5)	85+85+85=255	58.63%	68.26%	76.54%
Zoning+Centroid+Diagonal+ H_Peak (F1+F2+F3+F4)	85+170+85+85=425	74.13%	80.70%	88.26%
Zoning+Centroid+Diagonal+ V_Peak (F1+F2+F3+F5)	85+170+85+85=425	77.63%	81.46%	89.92%
Zoning+Centroid+H_Peak+ V_Peak (F1+F2+F4+F5)	85+170+85+85=425	61.76%	77.70%	84.56%

Features	Number of feature values	k-NN	Decision Tree	Random Forest
Zoning+Diagonal+H_Peak+ V_Peak (F1+F3+F4+F5)	85+85+85+85=340	70.03%	78.56%	84.16%
Centroid+Diagonal+H_Peak+ V_Peak (F2+F3+F4+F5)	170+85+85+85=425	58.83%	67.80%	74.28%
Zoning+Centroid+Diagonal+ H_Peak+V_Peak (F1+F2+F3+F4+F5)	85+170+85+85+85=510	81.57%	72.14%	79.26%

6.4 COMPARISON WITH THE EXISTING APPROACHES AND SYNTACTIC ANALYSIS

Based on the results, the comparative analysis of the present work with state-of-the-art work has been demonstrated in Table 6.3.

Table 6.3. Comparative analysis of the present work and state-of-the-art work

Authors	Script	Dataset	Feature Extraction Technique	Classification Technique	Accuracy
Gunter and Bunke (2004)	Latin	IAM	Geometric features	HMM with (i) Bagging (ii) Adaboost (iii) Random subspace (iv) Ensemble methods	66.23% (i) 67.92% (ii) 68.86% (iii) 68.67% (iv) 68.76%
Pal *et al.* (2011)	Bangla	4450 handwritten street name samples	Directional features	MQDF	91.13%
Pal *et al.* (2012)	Bangla, Devanagari and Latin	16,132 handwritten city name samples	Histogram of directional chain code features	MQDF	92.25%

Authors	**Script**	**Dataset**	**Feature Extraction Technique**	**Classification Technique**	**Accuracy**
Kumar *et al.* (2019)	Gurumukhi	1140 handwritten character samples	Zoning, DCT and gradient features	(i) k-NN (ii) SVM (iii) Decision Tree (iv) Random Forest	(Bagging) (i) 87.73% (ii) 92.19% (iii) 73.23% (iv) 87.36%
Present work	Gurumukhi	15,000 handwritten place name samples	Zoning, centroid, diagonal and peak extent features	Bagging methodology (i) k-NN (ii) Decision Tree (iii) Random Forest	(i) 84.96% (ii) 81.46% (iii) 89.92%

After a comparative analysis of the present approach with state-of-the-art approaches, the following key points have been analyzed:

- In Indian postal automation, work has been considered in the literature (Pal *et al.*, 2011; Pal *et al.*, 2012) based on MQDF with good accuracy, which is comparable with the present work.
- As presented in the literature (Gunter and Bunke, 2004) and the present experimental approach, there are different ensemble techniques available such as Adaptive Boosting and Bagging that yielded significant enhancements in accuracy than the base classifiers.
- For the Gurumukhi script, the attained accuracy via the present work is comparable with the work carried out by Kumar *et al.* (2019).
- Based on the Bagging methodology, there was a significant spike in the accuracy to 89.92% using the Random Forest classifier, which was superior as compared to the rate attained through the base classifier only.
- The results reveal the benefit of hybrid of the considered features in order to get spike in the accuracy. Based on the Bagging methodology, Random Forest

and Decision Tree classifiers attained their best performance on the hybrid of zoning, centroid, diagonal, and vertical peak extent features.

6.5 CHAPTER SUMMARY

In this chapter, the offline handwritten Gurumukhi word (place name) recognition system which finds its application in postal automation has been presented. The desirable features from the handwritten words were extracted based on four feature extraction techniques, namely, zoning features, centroid features, diagonal features and peak extent features (horizontal peak extent features and vertical peak extent features). For classification purpose, three classification techniques, namely, k-NN, Decision Tree, and Random Forest were employed. Out of the three considered classifiers, the Random Forest classifier attained a maximum accuracy of 88.86%. To attain a remarkable spike in recognition results, the Bagging methodology was utilized. Based on the Bagging methodology, the maximum accuracy of 89.92% was attained by employing zoning, centroid, diagonal, and the vertical peak extent features to the Random Forest classifier. Hence, the Bagging methodology proved fruitful to enhance the system performance.