Soft Computing (2021) 25:11589–11601
https://doi.org/10.1007/s00500-021-06060-1

DATA ANALYTICS AND MACHINE LEARNING

Improved recognition results of offline handwritten Gurumukhi characters using hybrid features and adaptive boosting

Munish Kumar[1] · M. K. Jindal[2] · R. K. Sharma[3] · Simpel Rani Jindal[4] · Harjeet Singh[5]

Accepted: 12 July 2021 / Published online: 25 July 2021
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract

Offline handwritten character recognition is a part of the arduous area of research in the domain of document analysis and recognition. In order to enhance the recognition results of offline handwritten Gurumukhi characters, the authors have applied hybrid features and adaptive boosting approach in this paper. On feature extraction stage, zoning, diagonal, centroid, and peak extent-based features have been taken into account for extracting the meaningful information about each character. On the classification stage, three classifiers, namely decision tree, random forest, and convolution neural network classifier, are used. For experimental work, the authors have collected 14,000 pre-segmented samples of Gurumukhi characters (35-class problem) written by 400 writers where they have used 70% data as training set and remaining 30% data as testing set. The authors have also explored fivefold cross-validation technique for experimental work. The AdaBoost approach along with the fivefold cross-validation strategy outstands the existing techniques in the relevant field with the recognition accuracy of 96.3%.

Keywords Character recognition · Classification · Decision tree · CNN · Random forest tree · Adaptive boosting

1 Introduction

Transferring data among human beings and computers play an important part in Document analysis and recognition. Optical character recognition (OCR) system is a key portion of a document analysis system and evolved for the identification of printed text as well as handwritten text because each document is segmented into lines, words, and characters for recognition and analysis of document text. Character recognition is a procedure which connects a figurative definition with objects (*letters*, *symbols,* and *numerals*) drawn on an image. There are twenty-two official languages and about twelve distinct scripts in India existing to compose these official Indian languages, and Gurumukhi is also one part of these scripts. For Punjabi language, Gurumukhi script is used for writing and is the twelfth most publicly spoken language in the world. Gurumukhi is composed of 35 basic characters in which there are 3 vowel bearers and 32 consonants. In addition to this, Gurumukhi script includes 6 additional consonants, 3 half characters and 9 vowel modifiers. The left to right and top to bottom is the pattern followed as the writing style in Gurumukhi script and is case insensitive. Even today, many algorithms have been put forward for character recognition, but the effectiveness of these algorithms is not satisfactory. So, there is still a scope to work on Gurumukhi character recognition and need to improve the recognition accuracy achieved with different methodologies. Moreover, handwritten Gurumukhi character recognition system finds applications in numerous fields like handwritten notes reading, form processing, recognition of handwritten postal code, etc. In this article, four feature

✉ Munish Kumar
munishcse@gmail.com

[1] Department of Computational Sciences, Maharaja Ranjit Singh Punjab Technical University, Bathinda, Punjab, India

[2] Department of Computer Science & Applications, Panjab University Regional Centre, Muktsar, Punjab, India

[3] Department of Computer Science & Engineering, Thapar Institute of Engineering & Technology, Patiala, Punjab, India

[4] Department of Computer Science & Engineering, Yadavindra College of Engineering, Talwandi Sabo, Bathinda, Punjab, India

[5] Chitkara University Institute of Engineering and Technology Chitkara University, Chandigarh, Punjab, India

extraction approaches, viz., zoning, peak extent-based, diagonal, and centroid features are evaluated. These feature extraction approaches have been evaluated by numerous researchers for distinct script identification, and it has been observed that these approaches are accomplishing recommendable results as compared to other approaches specified for handwriting recognition (Kumar et al. 2013a, b). Initially, each input image is partitioned into *n* number of uniform sized zones for extracting various features (Kumar et al. 2014). Depending on the four features specified, three classification techniques, namely, convolution neural network (CNN), decision tree and random forest are used for the classification of input characters. The authors have improved the recognition accuracy for offline handwritten Gurumukhi characters using adaptive boosting (AdaBoost) technique in the present work. This is an effort in direction to the establishment of the system that could recognize handwritten Gurumukhi characters effectively. The authors have noticed that most of the work for text recognition is carried out for non-Indic scripts and very few attempts have been made on Indian scripts. Various challenges have been identified like connected characters, the variation in the character patterns, uncertainty and indecipherability of character, overlapping characters in a word, unconstrained cursive words, variations in human writing styles, variations in shapes and sizes of character, poor input quality, and lesser accuracy rate in Gurumukhi character recognition. Other main considerations for Gurumukhi text recognition are the presence of distortions in headlines of the words, touching or overlapping characters, and different shape of a same character at different occurrences. Several methods have already been proposed for recognition of Gurumukhi character, but these are not able to achieve considerable recognition accuracy because quality of features and classifier is not able to achieve acceptable recognition accuracy. This motivated us to propose a hybrid approach of various techniques to improve the recognition accuracy for Gurumukhi character recognition. Moreover, the authors observed that there are numerous ensemble techniques (such as Bagging, Boosting) available in the literature that aid in attaining higher recognition rates as compared to using individual classifiers only. Although the concept of recognition in Gurumukhi script is pretty old, till now AdaBoost ensemble technique has not been employed for Gurumukhi character recognition. This fact motivated us to employ AdaBoost technique in order to attain superior performance in offline handwritten Gurumukhi character recognition system. The rest of the segments of this work are structured as defined. Section 2 details the related work specific to character recognition techniques. Section 3 discusses the various phases of the presented framework. Section 4 introduces AdaBoost methodology along with its working. Section 5 represents experimental results and the comparative analysis of performance of different classification techniques. The relative results of the presented work with the state-of-the-art work are reported in Sect. 6. Finally, in Sect. 7 the authors have concluded the paper and present a few future directions.

2 Related work

The recognition of offline handwritten Gurumukhi character is a difficult task as the consequence of different writing styles of individuals, various shapes of characters, and diverse sizes. A few writers use cursive writing, which also makes the character recognition system more complicated. A concise overview of related work on character recognition is presented in this section. Various researchers have dealt with handwritten character recognition for different scripts like Arabic, Bangla, Devanagari, etc. For example, Schwenk and Bengio (1997) applied the AdaBoost approach for decreasing the error rate for the recognition of online written characters, written by 203 individual writers which are classified by Diabolo networks and multilayer networks. Schomaker and Segers (1999) proposed a technique using geometrical features for the recognition of cursive Roman handwriting. Alaei et al. (2010) employed SVM classifier on modified chain code features for the recognition system of Persian isolated handwritten character which attains a recognition accuracy of 98.1%. Ahranjany et al. (2010) presented the combination of CNN with gradient training approach to recognize handwritten Farsi/Arabic digits by achieving the accuracy of 99.17%. Zhu et al. (2010) worked on 35, 686 samples of online handwritten Japanese text by employing a robust model by achieving a recognition accuracy of 92.8%. Rampalli and Ramakrishnan (2011) employed a combined recognition system for an online handwritten Kannada character recognition along with offline handwriting recognition which improves the accuracy of online handwriting recognizer by 11%. Venkatesh and Ramakrishnan (2011) recognized the online geometrical feature using a methodology that achieved an average accuracy of 92.6%. Gohell et al. (2015) worked on the recognition of online handwritten Gujarati characters and numerals using low level stroke feature-based technique, by attaining the numeral recognition accuracy of 95.0%, character recognition accuracy of 93.0% and combined numeral and character recognition accuracy of 90.0%. Saabni (2015) suggested the use of AdaBoost approach for the improvement in accuracy rates in the process of automatic handwriting recognition which has been tested on the 60,000 digits training samples. Antony et al. (2016) recognized the Tulu script by using the hybridization of Haar features and AdaBoost approach. Ardeshana et al. (2016) worked on the recognition of handwritten Gujarati character by extracting discrete cosine transformation-based features and using naïve Bayes classifier on 22,000 samples, thus achieving a recognition accuracy of 78.05%. Shahin (2017) proposed system for recognition of 14,000 different printed Arabic words using linear and nonlinear regression methodology which attains recognition accuracy of 86.0%. Dargan and Kumar (2018) surveyed the different feature extraction approaches and different classification methods used by numerous researches for writer identification in Indian and non-Indian scripts. Elbashir and Mustafa (2018) presented a CNN paradigm for the recognition of handwritten Arabic characters giving the test accuracy of 93.5% and training accuracy of 97.5% on the dataset presented by SUST ALT. Kumar et al. (2018a, b) explored a comprehensive review on character and numeral recognition of non-Indic and Indic scripts highlighting various constraints and issues faced. Gupta and Kumar (2019) attained the accuracy of 95.1% for the recognition of document forgery by extracting key printer noise feature, oriented FAST rotated and BRIEF feature, and speeded-up robust feature which is further classified by random forest classifier along with adaptive boosting approach. Husnain et al. (2019) recognized the unconstrained multi-font offline handwritten Urdu characters with an accuracy of 96.04% and 98.3% for Urdu characters and numerals, respectively, using CNN. Joseph et al. (2019) presented a model which is a combination of CNN as feature extraction tool and XGBoost as an accurate prediction model to increase the recognition rate handwritten text. This model is evaluated on 810,000 isolated handwritten English characters providing the accuracy of 97.18%. Kavitha and Srimathi (2019) used CNN to recognize 156 class offline handwritten isolated Tamil characters from the dataset of 82,929 images prepared by HP Labs India to produce the training and testing accuracy of 95.16% and 97.7%. Ptucha et al. (2019) presented an intelligent character recognition system based on the CNN approach recognizing random length handwritten text streams. Sethy and Patra (2019) achieved the average recognition accuracy of 98.8% on the implementation of Axis Constellation approach along with PCA on offline Odia handwritten characters. A few researchers have explored the area of handwritten Gurumukhi character recognition. For example, Garg (2009) exhibited a model for offline handwritten Gurumukhi character recognition, based on the idea of human biological neural network which normalizes input character size into 32 × 32 and extracts various features for character recognition. Kumar et al. (2011, 2012) have also presented a model in view of k-NN classifier for offline handwritten Gurumukhi character recognition achieving recognition accuracy of 94.12% for 3500 samples out of which only 350 samples have been considered for testing dataset. Siddharth et al. (2011) further published a model for offline handwritten Gurumukhi character recognition which explores support vector machine and probabilistic neural network classifiers. Singh and Budhiraja (2012) proposed a framework for Gurumukhi numeral recognition using wavelet transform features and backpropagation neural network for classifying the numerals based on the extracted features, accomplishing recognition accuracy of 88.83%. Also, a detailed study of various feature extraction approaches and classifiers and their combinations is presented. Koundal et al. (2017) studied the system for Punjabi printed and handwritten character recognition exploring its numerous feature extraction and classification approaches. Kaur and Kumar (2021) have proposed a system that uses a holistic approach to recognize a word, where a word itself is considered as an individual item. They employed various statistical features and different classifiers for the recognition purpose. They explored AdaBoost (adaptive boosting) algorithm to boost the recognition accuracy. Using AdaBoost, they attained maximum recognition accuracy of 88.78% for handwritten Gurumukhi word recognition. But, till date, researchers could not get acceptable recognition accuracy for offline handwritten Gurumukhi character recognition; therefore, there is still scoped to work on this to enhance the recognition accuracy. In this paper, the authors have captured 14,000 Gurumukhi character samples written by four hundred individuals and improved the recognition accuracy by using AdaBoost methodology. Yuan et al. (2019) proposed Gated CNN in order to integrate numerous convolutional layers for object detection. This proposed approach was experimented on two image datasets such as PASCAL VOC and COCO and attained promising results in terms of accuracy and speed as compared to the existing approaches. Masita et al. (2020) provided a review on the application of deep learning algorithms in object detection. They inferred that deep neural networks, convolutional neural networks, and region-based convolutional neural networks have been employed as the baseline for numerous robust detection systems. The deep learning proved to be efficient in detection of objects. Wang et al. (2021) proposed background and foreground seed selection approach in order to detect salient objects. They conducted the experiments on publicly available datasets and attained better performance than state-of-the-art approaches. In this work, the authors have considered noise-free dataset for the experimental work. But this methodology can be applicable for other documents and noise from those documents can be removed by using various techniques proposed by the Bhadouria et al. (2014). They have proposed various techniques and presented many studies to noise reduction techniques.

3 Description of the proposed system

The different phases used in the system for the character recognition of offline handwritten are digitization and pre-processing, feature extraction, and classification as shown in Fig. 1.

3.1 Digitization and pre-processing

Transforming the handwritten text written on paper to a computerized form is known as digitization. This computerized conversion is completed by scanning the document, and afterward the bitmap image '.bmp' is created of the original scanned text document. For online character recognition, the numbers of strokes are available, whereas for offline character recognition, the document is scanned, and one can get the digital image of that document. The digital image is produced in the digitization process which becomes the input for the pre-processing phase, which is the preliminary phase in the presented system. In this paper, authors produced the digital images of paper-based text documents using the HP-1400 scanner. In pre-processing phase, the digitized input images by using nearest neighborhood interpolation (NNI) algorithm are normalized in the size of 88 × 88 pixels, and also, the authors had diminished the line width of the text using the thinning approach, viz., parallel thinning algorithm proposed by Zhang and Suen (1984).

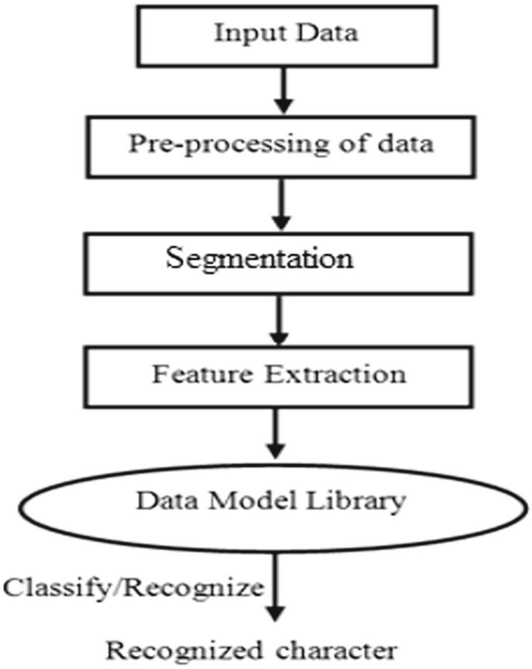

Fig. 1 Sequence of phases for handwritten character recognition system

3.2 Feature extraction

It is used to extricate the relevant details contained in the digital image of the input character. Performance of optical character recognition system is basically relying on the extracted details to great extent. Various feature extraction techniques have been found in the literature, which are used in the character recognition system to recognize the characters. In optical character recognition applications, it is critical to extract those features that allow the system to differentiate between all the character classes that exist. Feature extraction approaches are classified into two types, namely structural and statistical features. In the present work, the authors have extracted statistical features, viz., zoning, peak extent-based, centroid, and diagonal features. The zoning feature extraction technique divides the character image into *n* number of regions in hierarchical arrangement as proposed by Kumar et al. (2014). Suppose an input image is at current level L and then it is having $4^{(L)}$ sub-images. For instance, the number of sub-images is 4 for $L = 1$, 16 for $L = 2$, and so on. Therefore, $4^{(L)}$ -dimensional feature vector is produced for every L. In this work, $L = [0, 1, 2, \text{and } 3]$ has been considered by the authors.

In the diagonal feature extraction methodology, initially a character image is divided into *n* number of regions in hierarchical order. For each region, by moving along the diagonals of its respective pixels, the features are extracted. The center of an image is the point which is considered as the centroid which becomes the feature by considering the coordinates of the centroid of the foreground pixels in each region of an input image. By considering the summation of the peak extents which is calculated with the consecutive black pixels among each region of an input image leads to the formation of the peak extent-based feature.

The above-specified feature extraction approaches have been implemented in various script recognition systems by different researchers, which concludes the better performance of these techniques in comparison with other approaches defined for handwriting recognition systems. Also, the performance evaluation of peak extent-based features is evaluated better by Kumar et al. (2013a, 2015) among the other existing approaches for handwritten Gurumukhi. So, in this presented work, for enhancing the recognition accuracy of handwritten Gurumukhi character recognition system, these feature extraction techniques individually and blend of these four feature extraction techniques in the hybrid structure has been implemented.

Algorithm (a): Zoning Feature Extraction.

To extract these features, the following steps have been used:

Step I The thinned image is divided up to four levels in hierarchical order to create zones.

Step II At each level and for each zone, the number of foreground pixels is calculated.

The above steps, for each L level, will provide a feature set with $4^{(L)}$elements.

Algorithm (b): Diagonal Feature Extraction.

For diagonal feature extraction methodology, following steps is used:

Step I The thinned image is divided up to four levels in hierarchical order to create zones.

Step II For a single sub-feature in each zone, diagonal foreground pixels present are summed.

Step III Then, to create a single value feature for the corresponding zone, the values calculated in Step II are averaged.

Step IV For the zones which do not have any diagonal foreground pixel, its feature value is considered as zero.

The above steps, for each L level, will provide a feature set with $4^{(L)}$ elements.

Algorithm (c): Centroid Feature Extraction.

Step I The thinned image is divided up to four levels in hierarchical order to create zones.

Step II At each level and for each zone, calculate the coordinates of foreground pixels.

Step III For the feature value, store the coordinates of the calculated centroid from the values taken from Step II.

Step IV The feature value is considered as zero for the zones having no foreground pixel.

The above steps, for each L level, will generate a feature set with $2 \times 4^{(L)}$ elements.

Algorithm (d): Peak Extent Feature Extraction.

Step I The thinned image is divided up to four levels in hierarchical order to create zones.

Step II For each row in a zone, calculate the sum of successive foreground pixels as the peak extent.

Step III For each row in a zone, the values of successive foreground pixels are replaced by peak extent value.

Step IV In each row, find the largest value of peak extent.

Step V For the respective zone, the summation of the values of peak extent sub-feature is considered as feature.

Step VI The feature value is considered as zero for the zones having no foreground pixel.

Step VII The values in the feature vector are normalized by dividing each element of the feature vector with the largest value in the feature vector.

The above steps, for each L level, will generate a feature set with $2 \times 4^{(L)}$ elements.

3.3 Classification

Classification is also considered as the crucial functions in the character recognition process. It is the last phase of a character recognition process in which unique labels are assigned based on the extracted features to character images. Classification basically selects the feature space to which the unknown pattern belongs, and by using different analytical properties and artificial intelligence approaches, it extracts samples from a few classes. In this process, the feature vectors of the input character are compared with the illustration of each character class with the main aim to reduce the misclassification admissible to the features extracted in the previous state. In this work, authors have considered three classification methodologies, i.e., random forest, convolution neural network, and decision tree. These classifiers are most widely used these days in the fields of image processing and pattern recognition with CNN as the most appropriate for pattern recognition. And these classifiers are giving good performance for text recognition of various non-Indic and Indic scripts. So, these three classifiers are considered in this work for handwritten Gurumukhi character recognition. The authors have used LeNet of CNN for character recognition (LeCun et al. 1998). A decision tree classifier is communicated as a recursive part of occurrence space. The decision tree is made up of nodes that give a rooted tree, *i.e.,* a node termed as "root" with a property of no incoming edges forms a directed tree and absolutely one incoming edge is present for all other nodes. The representation used for the decision tree is a test on an attribute (feature) is denoted by each internal node, the outcome of the test is represented on each branch of the tree, and the value of the class-label is given by each leaf (terminal or decision node). The overfitting critical situation of decision tree is eliminated using random forest. The main usage of decision tree classifiers is to classify the numerous sub-samples of the dataset, and for such a design, the meta estimator that fits this statistic of decision tree classifiers is achieved with the use of random forest. So, among the supervised learning classification algorithms and for large databases, random forest achieves unexcelled (Breiman 2001). Experimental results using these classification techniques are presented in Sect. 5.

4 AdaBoost methodology

AdaBoost introduces to a specific training approach for a boosted classifier. A boost classifier is defined as: -

$$F_T(x) = \sum_{t=1}^{T} f_t(x)$$

where an object x is taken as input by each weak learner f_t and a value is returned as output expressing the class of the object. Adaptive boosting (AdaBoost), proposed by Freund and Schapire (1999) classified as machine learning meta-algorithm. Adaptive boosting is a representative algorithm in the boosting family. The basic idea of this algorithm is to be used in collaboration with other learning algorithms to enhance their performance. AdaBoost is robust in terms of consecutive weak learners, which are twisted for those features which are misclassified by earlier classifiers. The AdaBoost algorithm is the most widely used algorithm in various applications of artificial intelligence and pattern recognition. AdaBoost is a kind of classifier with huge recognition accuracy. It is straightforward in establishment and does not have to make selections of high-quality features using PCA and so forth. In this paper, the authors have considered this approach for enhancing the recognition results of offline handwritten Gurumukhi characters. In this present work, three classifiers, namely random forest, convolution neural network and decision tree, are evaluated as mentioned above in Sect. 4. Convolution neural network performs better than other two classifiers in this work, and using the convolution neural network, a recognition accuracy of 90.4% was accomplished. Therefore, finally authors have applied the AdaBoost algorithm to convolution neural network for improving the recognition results and obtained a recognition accuracy of 96.3%. Moreover, root mean squared error and false acceptance rates of the proposed framework have also been reduced with AdaBoost algorithm. The experiential outcomes based on three classifiers and AdaBoost algorithm as specified above are discussed in the section below.

5 Experimental results and discussions

Four feature extraction methodologies, namely peak extent-based, diagonal, centroid and zoning features, have been used to extract the shape information of characters for recognition. A combination of these features in a hybrid manner has also been tried for enhancing the recognition accuracy. In order to recognize offline handwritten Gurumukhi characters, three classifiers, namely random forest, decision tree and convolution neural network, have been considered in this work along with 35 basic characters (three vowel bearers and thirty-two consonants) of the Gurumukhi script. Authors have collected 14, 000 samples of offline handwritten Gurumukhi characters on which each approach has been tested. These samples were collected from four hundred writers at public places like, schools, colleges etc. The dataset is partitioned into two portions using fivefold cross-validation technique with 70% data of 14,000 samples is considered as training dataset, whereas the remaining samples (30% data) are considered as testing dataset for performance analysis. The results of various experiments show that our algorithm can accomplish a higher recognition rate (up to 96.3%), and a little higher than the convolution neural network using fivefold cross-validation technique, when we used CNN with adaptive boosting methodology. Classifier-wise recognition results based on partitioning strategy and fivefold cross-validation technique are depicted in Tables 1, 2, respectively. These results are graphically depicted in Figs. 2, 3, respectively. Moreover, the root mean squared error (RMSE) as 5.20% and the false acceptance rate (FAR) as 1.0% are also reduced using the AdaBoost methodology as depicted in Tables 3, 4, 5, 6. RMSE is graphically presented for partitioning strategy and fivefold cross-validation technique in Figs. 4, 5, respectively, and similarly, FAR for partitioning strategy and fivefold cross-validation technique in Figs. 6, 7, respectively. The authors have also noticed that hybrid features seem like the most discriminating features as compared to other features as shown in Table 7.

6 Comparison with state-of-the-art work

In this segment, authors have presented comparisons of proposed methodology with the extant work. The recognition accuracy of 91.6% has been attained by Lehal and Singh (1999) for printed Gurumukhi text recognition. For handwritten Gurumukhi character recognition, Sharma and Jain (2010) experimented on extricated zoning features and two classifiers, namely k-NN and SVM, achieving maximum 72.5% and 72.0% recognition accuracy with k-NN and SVM, respectively. Kumar et al. (2011) have experimented 3500 samples of handwritten characters by extracting intersection and open end points-based features which have been classified by SVM classifier. The diagonal feature is also extracted achieving accuracy of 94.1% on the same sample set using k-NN classifier. It has been observed during experimentation that the number of training samples and testing samples in the dataset affects recognition accuracy.

Kumar et al. (2013b) attained accuracy of 84.6%, 85.9%, and 89.2% by experimenting on 10,500 offline handwritten Gurumukhi character samples using modified division points-based features, which are classified by Linear-SVM, MLP, and k-NN classifiers. Kumar et al. (2016) have used various transformation techniques for offline handwritten Gurumukhi character recognition and have accomplished a recognition accuracy of 95.8% for

Table 1 Recognition accuracy using partitioning strategy (70% data as training set and 30% data as testing set)

Feature extraction technique	Classification technique			
	CNN (%)	Decision tree (%)	Random forest tree (%)	AdaBoost_CNN (%)
Zoning features	89.17	61.46	82.22	89.86
Diagonal features	74.69	57.20	72.43	78.25
Centroid features	79.20	56.72	77.64	81.27
Peak extent features	71.26	56.22	70.29	75.90
Hybrid features	89.92	60.45	80.32	90.37

Table 2 Recognition accuracy using fivefold cross-validation

Feature extraction technique	Classification technique			
	CNN (%)	Decision tree (%)	Random forest tree (%)	AdaBoost_CNN (%)
Zoning features	89.90	61.70	83.90	92.10
Diagonal features	76.70	58.40	83.50	79.40
Centroid features	82.10	58.10	83.40	82.30
Peak extent features	71.70	58.10	76.80	75.90
Hybrid features	90.40	61.70	84.40	96.30

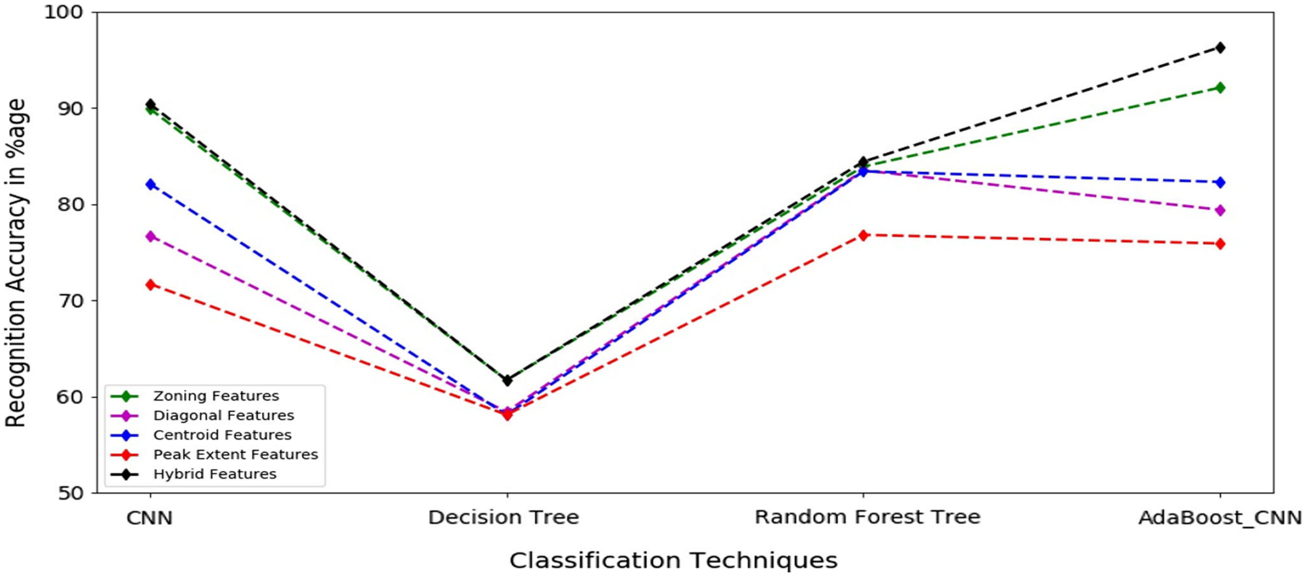

Fig. 2 Recognition accuracy using partitioning strategy

10,500 samples penned by 200 distinct writers. Kumar et al. (2018b) improved the recognition rate to 95.91% for the Medieval handwritten Gurumukhi 150 manuscripts by using the combination of zoning, DCT, and gradient features and combination of k-NN, SVM, decision tree, and random forest classifiers along with the application of adaptive boosting and bagging approach. Kumar et al. (2019) evaluated the performance of the various classifiers, viz., k-NN, SVM (Linear, RBF), naïve Bayes, decision tree, CNN, and random forest applied on the peak extent diagonal and centroid features extracted from the sample set of 13,000 Gurumukhi isolated characters and numerals. In this paper, the authors have considered 14,000 samples

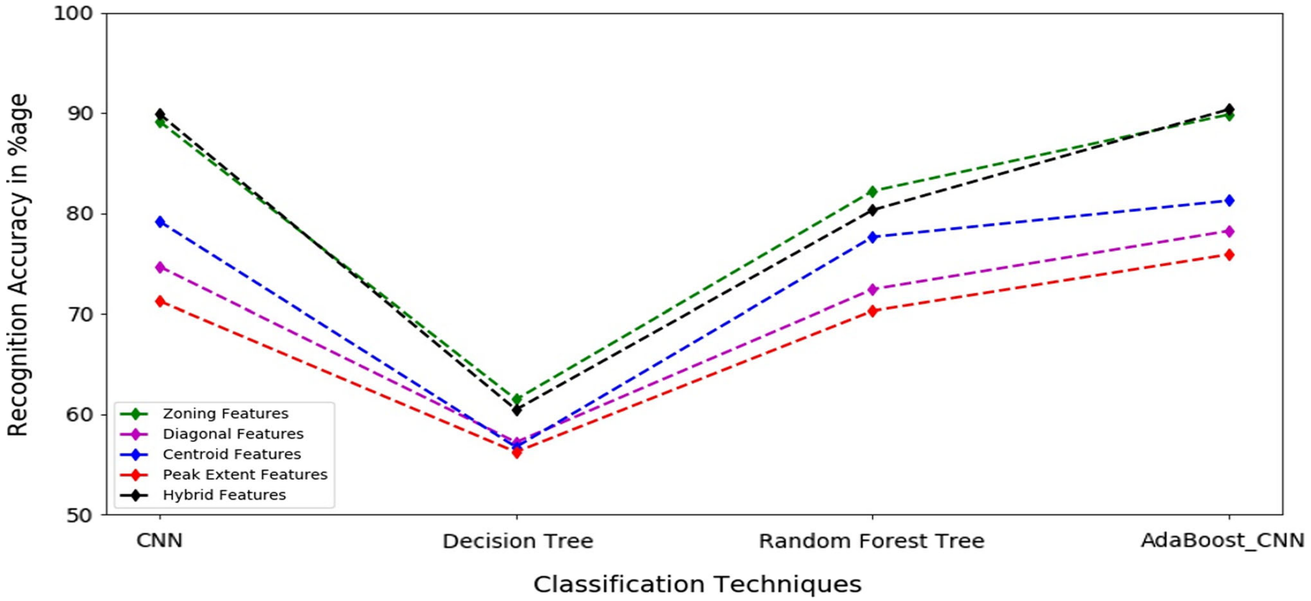

Fig. 3 Recognition accuracy using fivefold cross-validation

Table 3 RMSE using partitioning strategy (70% data as training set and 30% data as testing set)

Feature extraction technique	Classification technique			
	CNN (%)	Decision tree (%)	Random forest tree (%)	AdaBoost_CNN (%)
Zoning features	5.60	11.10	8.70	5.20
Diagonal features	8.50	11.60	8.70	8.10
Centroid features	7.70	11.40	8.80	7.20
Peak extent features	9.50	11.60	9.40	8.80
Hybrid features	9.90	11.10	8.70	5.20

Table 4 RMSE using fivefold cross-validation

Feature extraction technique	Classification technique			
	CNN (%)	Decision tree (%)	Random forest tree (%)	AdaBoost_CNN (%)
Zoning features	5.71	11.32	8.87	5.32
Diagonal features	8.67	11.83	8.82	8.27
Centroid features	7.85	11.64	8.97	7.37
Peak extent features	9.69	11.73	9.56	8.97
Hybrid features	10.09	11.42	8.92	5.34

and obtained recognition accuracy of 96.3% using CNN and adaptive boosting methodology.

7 Conclusion and future scope

In this paper, the authors presented the improved recognition results for offline handwritten Gurumukhi character recognition system based on AdaBoost ensemble technique. Four features, namely peak extent-based, diagonal, centroid and zoning features, along with their combinations

Table 5 FAR using partitioning strategy (70% data as training set and 30% data as testing set)

Feature extraction technique	Classification technique			
	CNN (%)	Decision tree (%)	Random forest tree (%)	AdaBoost_CNN (%)
Zoning features	2.0	8.0	4.0	2.0
Diagonal features	5.0	7.0	4.0	3.0
Centroid features	4.0	8.0	3.0	3.0
Peak extent features	4.0	9.0	5.0	4.0
Hybrid features	3.0	6.0	4.0	2.0

Table 6 FAR using fivefold cross-validation

Feature extraction technique	Classification technique			
	CNN (%)	Decision tree (%)	Random forest tree (%)	AdaBoost_CNN (%)
Zoning features	3.0	7.0	3.0	2.0
Diagonal features	4.0	8.0	3.0	4.0
Centroid features	3.0	8.0	3.0	3.0
Peak extent features	5.0	8.0	4.0	4.0
Hybrid features	2.0	6.0	2.0	1.0

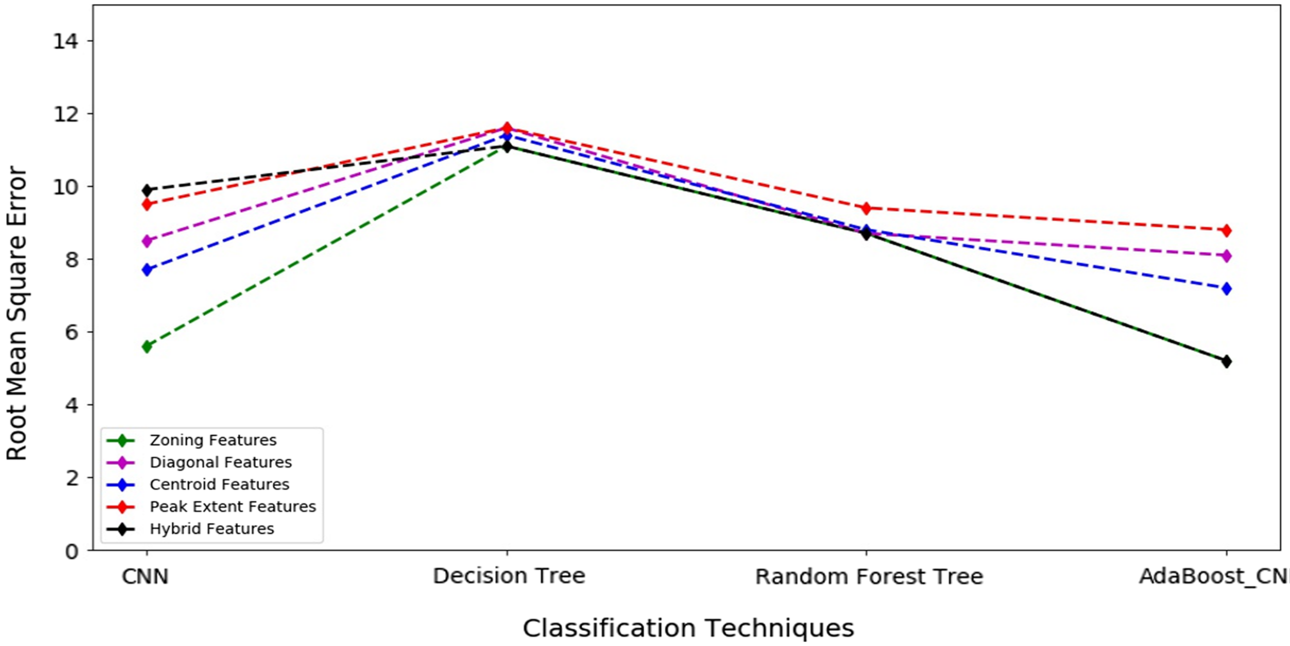

Fig. 4 RMSE using partitioning strategy

have been considered in order to extract the desired features from the character samples. The classifiers that have been employed in this work are convolution neural network, decision tree and random forest tree. Experimental results present the effectiveness of the proposed system based on recognition accuracy, RMSE and FAR. Using the considered classification techniques and fivefold cross-validation technique, the authors have accomplished a

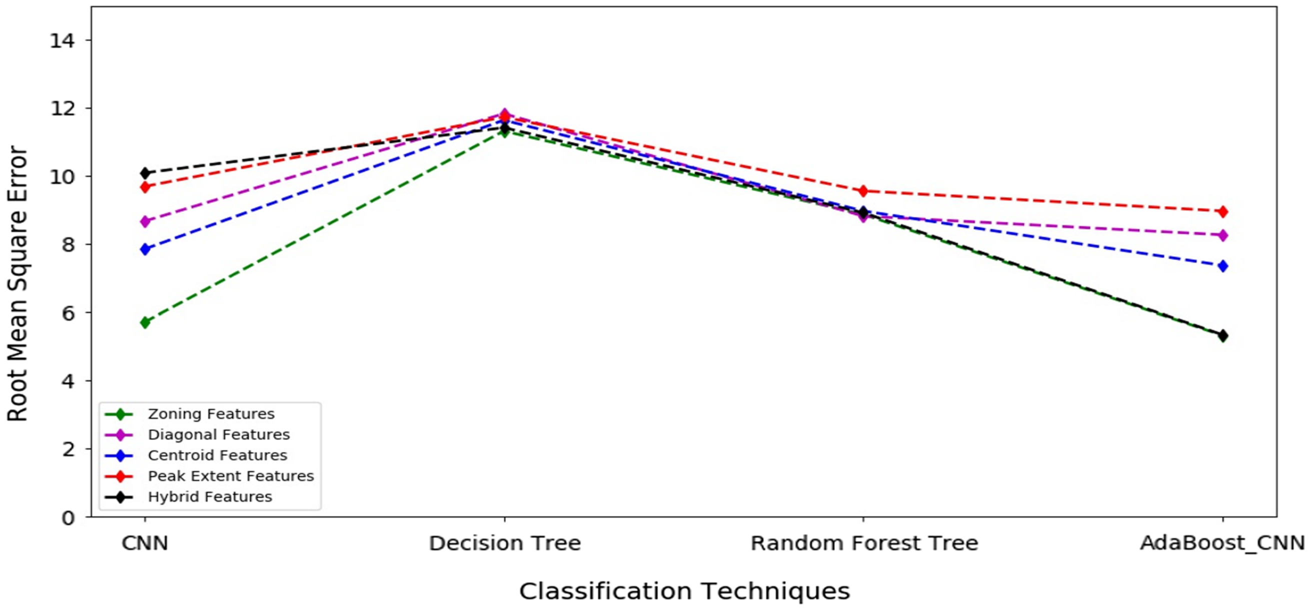

Fig. 5 RMSE using fivefold cross-validation

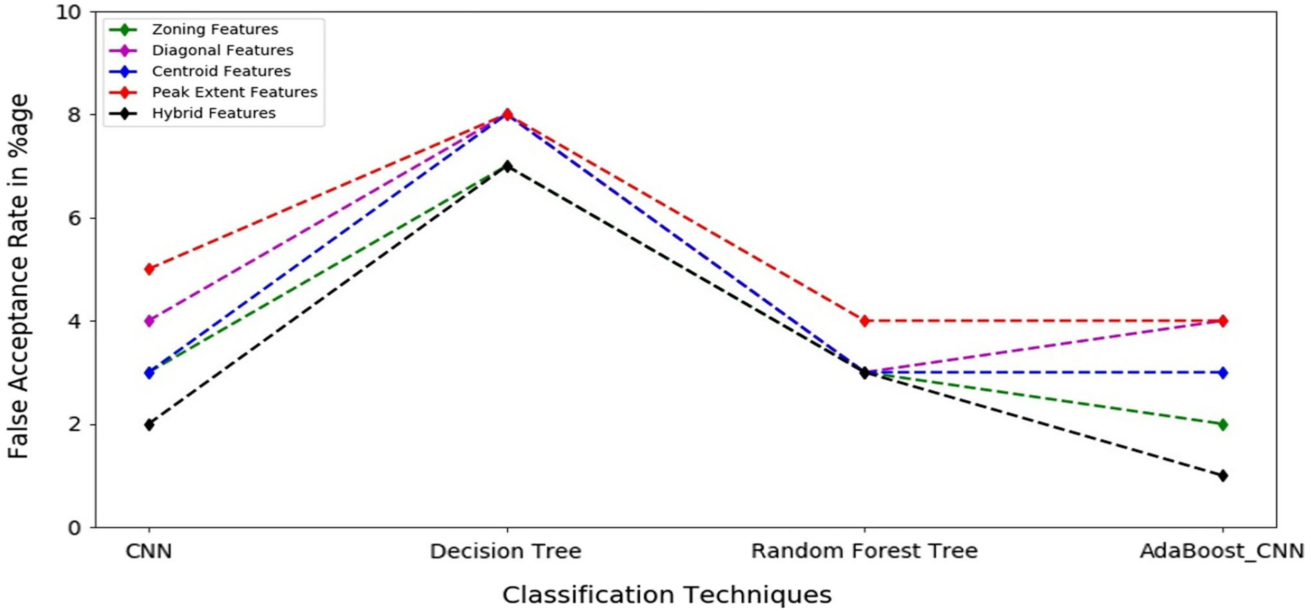

Fig. 6 False acceptance rate using partitioning strategy

recognition accuracy of 90.4%, 61.7% and 84.4% with CNN, decision trees and random forest tree, respectively, for 14,000 samples of training and testing dataset. Finally, the highest recognition accuracy of 96.3% has been attained using AdaBoost methodology and demonstrated the superior performance of the proposed approach as compared to other state-of-the-art approaches. To the best of authors' knowledge, AdaBoost technique has been applied in order to recognize offline handwritten Gurumukhi characters and thus the results attained can be considered as baseline for comparisons in the further research. This is well known that Gurmukhi shares a number of structural similarities with some other Indian scripts. As such, the work carried out in this thesis can further be extended for these scripts after building a database for these scripts.

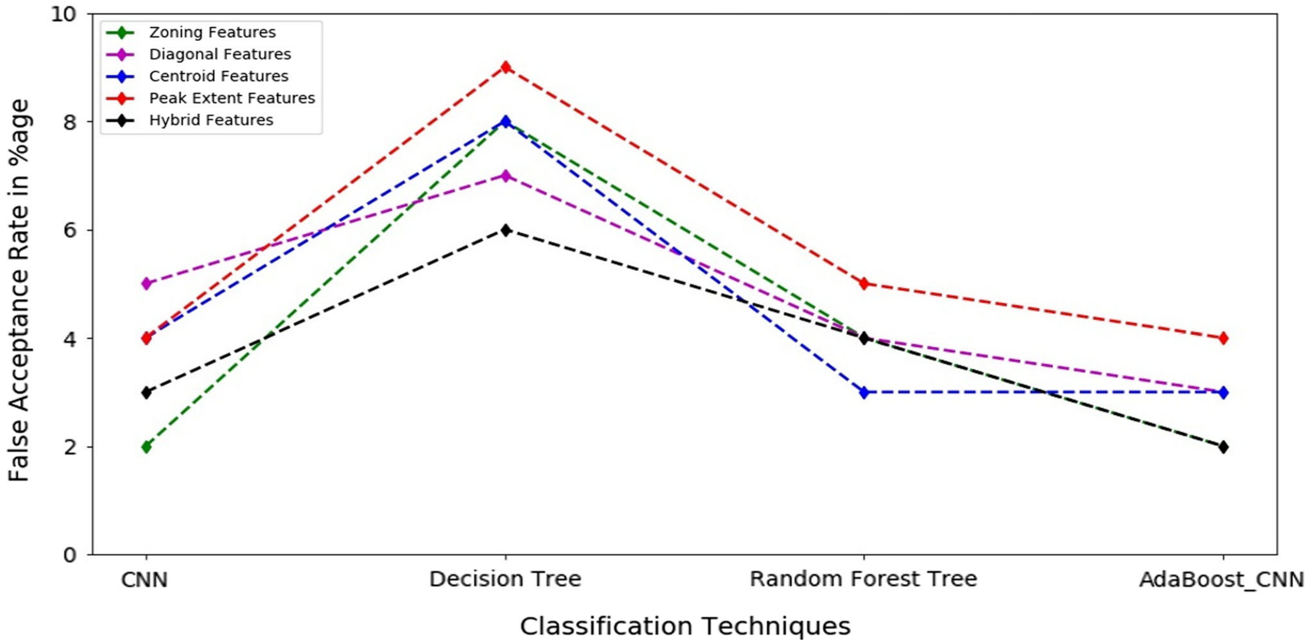

Fig. 7 False acceptance rate using fivefold cross-validation

Table 7 Comparison with existing methodologies for Gurumukhi character recognition

Authors	Feature extraction technique	Classification	Data set	Recognition accuracy
Lehal and Singh (1999)	Local and global features	k-NN	3000	91.6%
Sharma and Jain (2010)	Zoning	k-NN and SVM	15,000	72.5% (k-NN) and 72.0% (SVM)
Kumar et al. (2011)	Diagonal	k-NN	3500	94.1%
Kumar et al. (2013b)	Modified division points	Linear-SVM, MLP, k-NN	10,500	84.6% (Linear-SVM), 85.9% (MLP) and 89.2% (k-NN)
Kumar et al. (2016)	Discrete transformations	Linear-SVM	10,500	95.8%
Kumar et al. (2018b)	Zoning, Discrete cosine transformations, Gradient features	k-NN, SVM, decision tree, random forest	1140	95.91%
Kumar et al. (2019)	Peak Extent features, Diagonal, Centroid	k-NN, SVM (Linear, RBF), naïve Bayes, decision tree, CNN, random forest	13,000	87.9%
Proposed work	Hybrid features	CNN	14,000	96.3%

Declarations

Conflict of interest Authors have no conflict of interest.

References

Ahranjany SS, Razzazi F, Ghassemian MH (2010) A very high accuracy handwritten character recognition system for Farsi/Arabic digits using convolutional neural networks. In: Proceedings of the IEEE 5th international conference on bio-inspired computing: theories and applications (BIC-TA), pp. 1585–1592

Alaei A, Nagabhushan P, Pal U (2010) A new two-stage scheme for the recognition of Persian handwritten characters. In:

Proceedings of the 12th international conference on frontiers in handwriting recognition (ICFHR), pp. 130–135

Antony PJ, Savitha CK and Ujwal UJ (2016) Haar features based handwritten character recognition system for tulu script. In: Proceedings of the IEEE International conference on recent trends in electronics information communication technology, pp. 65–68

Ardeshana M, Sharma AK, Adhyaru DM, Zaveri TH (2016) Handwritten Gujarati character recognition based on discrete cosine transform. In: Proceedings of the IRF-IEEE forum international conference, pp. 23–26

Bhadouria VS, Ghoshal D, Siddiqi AH (2014) A new approach for high density saturated impulse noise removal using decision-based coupled window median filter. SIViP 8:71–84

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Dargan S, Kumar M (2018) Writer identification system for indic and non-indic scripts: state-of-the-art survey. Arch Comput Methods Eng. https://doi.org/10.1007/s11831-018-9278-z

Elbashir MK, Mustafa ME (2018) Convolutional neural network model for arabic handwritten characters recognition. Int J Adv Res Comput Commun Eng 7(11):1–5

Freund Y, Schapire RE (1999) A short introduction to boosting. J Jpn Soc Artif Intell 14(5):771–780

Garg N (2009) Handwritten Gurumukhi Character recognition using neural networks, M.E. thesis, Thapar University, Patiala, India

Gohell CC, Goswam MM, Prajapate YK (2015) On-line handwritten Gujarati character recognition using low level stroke. In: Proceedings of the third international conference on image information processing, pp. 130–134

Gupta S, Kumar M (2019) Forensic document examination system using boosting and bagging methodologies. Soft Comput. https://doi.org/10.1007/s00500-019-04297-5

Husnain M, Missen MMS, Mumtaz S, Jhandir MZ, Coustaty M, Luqman MM, Ogier JM, Choi GS (2019) Recognition of Urdu handwritten characters using convolutional neural network. Appl Sci. https://doi.org/10.3390/app9132758

Joseph JS, LakshmiKiranParthiban CUP (2019) An efficient offline handwritten character recognition using CNN and Xgboost. Int J Innov Technol Explor Eng 8(6):115–118

Kaur H, Kumar M (2021) On the recognition of offline handwritten word using holistic approach and AdaBoost methodology. Multimed Tools Appl 80:11155–11175

Kavitha BR, Srimathi C (2019) Benchmarking on offline handwritten Tamil character recognition using convolutional neural networks. J King Saud Univ-Comput Inf Sci. https://doi.org/10.1016/j.jksuci.2019.06.004

Koundal K, Kumar M, Garg NK (2017) Punjabi optical character recognition: a survey. Indian J Sci Technol 10(19):1–8

Kumar M, Sharma RK, Jindal MK (2013a) A novel feature extraction technique for offline handwritten Gurumukhi character recognition. IETE J Res 59(6):687–692

Kumar M, Jindal MK, Sharma RK (2013b) MDP Feature extraction technique for offline handwritten Gurumukhi character recognition. Smart Comput Rev 3(6):397–404

Kumar M, Sharma RK, Jindal MK (2014) A novel hierarchical technique for offline handwritten Gurumukhi character recognition. Natl Acad Sci Lett 37(6):567–572

Kumar M, Jindal MK, Sharma RK (2016) Offline handwritten Gurumukhi character recognition: analytical study of different transformations. Proc Natl Acad Sci, India, Sect A 87(1):137–143

Kumar M, Jindal MK, Sharma RK, Jindal SR (2018a) Character and numeral recognition for non-Indic and Indic scripts: a survey. Artif Intell Rev. https://doi.org/10.1007/s10462-017-9607-x

Kumar M, Jindal SR, Jindal MK, Lehal GS (2018b) Improved recognition results of medieval handwritten Gurumukhi manuscripts using boosting and bagging methodologies. Neural Process Lett. https://doi.org/10.1007/s11063-018-9913-6

Kumar M, Jindal MK, Sharma RK, Jindal SR (2019) Performance evaluation of classifiers for the recognition of offline handwritten Gurumukhi characters and numerals: a study. Artif Intell Rev. https://doi.org/10.1007/s10462-019-09727-2

Kumar M, Jindal MK, Sharma RK (2011) k-Nearest neighbor based offline handwritten Gurumukhi character recognition. In: Proceedings of the international conference on image information processing, Jaypee University of Information Technology, Waknaghat (Shimla), pp. 1–4

Kumar M, Jindal MK, Sharma RK (2012) Offline handwritten Gurumukhi character recognition: study of different features and classifiers combinations. In: Proceedings of the workshop on document analysis and recognition, IIT Bombay, pp. 94–99

Kumar M (2015) Offline handwritten Gurumukhi script recognition, Ph. D. thesis, Thapar University, Patiala, India

LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–2324

Lehal GS, Singh C (1999) Feature extraction and classification for OCR of Gurumukhi script. Vivek 12(2):2–12

Masita KL, Hasan AN and Shongwe T (2020) Deep learning in object detection: a review. In: Proceedings of international conference on artificial intelligence, big data, computing and data communication systems, pp. 1–11

Ptucha R, Such FP, Pillai S, Brockler F, Singh V, Hutowski P (2019) Intelligent character recognition using fully convolutional neural networks. Pattern Recogn 88(2019):604–613

Rampalli R, Ramakrishnan AG (2011) Fusion of complementary online and offline strategies for recognition of handwritten Kannada characters. J Univ Comput Sci (JUCS) 17(1):81–93

Saabni R (2015) Ada-boosting extreme learning machines for handwritten digit and digit strings recognition. In: Proceedings of the 5th international conference on digital information processing and communications.

Schomaker L, Segers E (1999) Finding features used in the human reading of cursive handwriting. IJDAR 2:13–18

Schwenk H, Bengio Y (1997) AdaBoosting neural networks: application to on-line character recognition. In: Proceedings of the international conference on artificial neural network, pp. 967–972

Sethy A, Patra PK (2019) Off-line Odia handwritten character recognition: an axis constellation model based research. Int J Innov Technol Explor Eng 8(9S2):788–793

Shahin AA (2017) Printed Arabic text recognition using linear and non-linear regression. Int J Adv Comput Syst Appl 8(1):227–235

Sharma DV, Jain U (2010) Recognition of isolated handwritten characters of Gurumukhi script using neocognitron. Int J Comput Appl 10(8):10–16

Siddharth KS, Jangid M, Dhir R, Rani R (2011) Handwritten Gurumukhi character recognition using statistical and background directional distribution features. Int J Comput Sci Eng 3(6):2332–2345

Singh P, Budhiraja S (2012) Offline handwritten Gurumukhi numeral recognition using wavelet transforms. Int J Mod Educ Comput Sci 8:34–39

Venkatesh N, Ramakrishnan AG (2011) Choice of classifiers in hierarchical recognition of online handwritten Kannada and Tamil aksharas. J Univ Comput Sci (JUCS) 17:94–106

Wang H, Zhu C, Shen J, Zhang Z, Shi X (2021) Salient object detection by robust foreground and background seed selection. Comput Electr Eng. https://doi.org/10.1016/j.compeleceng.2021.106993

Yuan J, Xiong H-C, Xiao Y, Guan W, Wang M, Hong R, Li Z-Y (2019) Gated CNN: integrating multi-scale feature layers for object detection. Pattern Recogn 105:1–28

Zhang TY, Suen CY (1984) A fast parallel algorithm for thinning digital patterns. Commun ACM 27(3):236–239

Zhu B, Zhou XD, Liu CL, Nakagawa M (2010) A robust model for on-line handwritten Japanese text recognition. Int J Doc Anal Recognit (IJDAR) 13(2):121–131

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.