Writer Identification System for Handwritten Gurmukhi Characters: Study of Different Feature-Classifier Combinations

Sakshi, Naresh Kumar Garg and Munish Kumar

Abstract In this paper, we are exploring various features and classifiers for writer identification in light of Gurmukhi text handwriting. The identification of the writers based on a piece of handwriting is a challenging task for pattern recognition. The writer identification framework proposed in this paper includes diverse stages like image preprocessing, feature extraction, training, and classification. The framework first prepares a skeleton of the character so that meaningful data about the handwriting of writers can be extracted. The feature extraction stage incorporates various plans, namely, zoning, diagonal, transition, intersection and open end points, centroid, the horizontal peak extent, the vertical peak extent, parabola curve fitting, and power curve fitting based features. In order to assess the prominence of these features, we have used four classification techniques, namely, Naive Bayes, Decision Tree, Random Forest and AdaBoostM1. For experimental results, we have collected 49,000 samples from 70 different writers. In this work, maximum accuracy of 81.75% has been obtained with centroid features and AdaBoostM1 classifier.

Keywords Feature extraction · Classification · Naive bayes · Decision tree
Random forest · AdaBoostM1

Sakshi · N.K. Garg
Department of Computer Science & Engineering, GZS Campus College
of Engineering & Technology (Maharaja Ranjit Singh Punjab Technical University),
Bathinda, Punjab, India
e-mail: me.sakshi51@gmail.com

N.K. Garg
e-mail: naresh2834@rediffmail.com

M. Kumar (✉)
Department of Computer Applications, GZS Campus College of Engineering & Technology
(Maharaja Ranjit Singh Punjab Technical University), Bathinda, Punjab, India
e-mail: munishcse@gmail.com

N. Chaki et al. (eds.), *Proceedings of International Conference on Computational Intelligence and Data Engineering*, Lecture Notes on Data Engineering
and Communications Technologies 9, https://doi.org/10.1007/978-981-10-6319-0_11

1 Introduction

Writing is implied by the representation of language in a textual medium using an arrangement of 'signs or symbols'. Every individual has his own writing style which relies on a considerable measure of variables like particular shape of letters, spacing between letters, etc. Handwriting of a person is also subject to the mental condition of the individual like his level of inspiration, anger, joy, and others. In any case, it is found that handwriting of a person is generally steady however might be influenced gradually with age. Writer identification, in general plays an important role in forensic, writer verification schemes, and related branches of science. The parameters are typically considered for writer identification is comprehensiveness uniqueness, aging, accessibility, etc. The same is applicable in Gurmukhi script too with an extra essentialness of finding the advancement and development of the Gurmukhi script. It is clear that the significance of writer identification has turned out to be more vital nowadays. So, the number of researchers engaged with this challenging issue, is going on top of these opportunities. There are many languages all through the world. Each language represents an alternate danger to the writer identification issue contingent upon the characteristics of the language. It is clear that the identification problem differs across various languages. The handwriting-based writer identification is an active research area. A comprehensive survey of writer identification till 1989 was presented by Lorette and Plamondon [8, 11]. Zois and Anastassopoulos [14] have presented a writer identification framework based on English and Greek writers. They have achieved an accuracy of 95% for writer identification of both English and Greek writers. Leeham et al. [9] proposed a technique to recognize the writers based on numerals written by them. Schlapbach and Bunke [12, 13] presented a Hidden Markov Model (HMM) based writer identification and verification strategy. They prepared an individual HMM for each writer's handwriting. The identification technique was tested with data set collected from 650 scholars. Gazzah and Amara [2] have proposed an approach for writer identification framework based on offline handwritten Arabic script documents. The proposed technique depends on combining the global and neighborhood feature sets by using the genetic algorithm so as to take out the redundant and irrelevant features. They have considered two classifiers, SVM and MLP for recognition. They have noticed that MLP performs preferable outcomes over SVM and got precision of around 94%. Ghiasi and Safabakhsh [3] have proposed an effective technique for writer recognition using a code book. They have utilized Farsi database which incorporates short, medium, and extensive messages and results demonstrate that the productivity of short messages are more effective. Maadeed [10] has proposed a writer identification framework based on Arabic handwritten text. He has also compared the edge direction distribution features with other features of Arabic text and used *k*-NN classifier for recognition. Writer identification framework is divided into two categories, namely, text dependent and text independent. Depending on the text content, text-dependent methods match the same characters and hence require the writer to write the same text. Text-independent methods are able to identify writers independent of the text content and it does not require comparison of the same characters. A very few studies in Indian languages have been documented so far. Currently, writer identification of handwritten Gurmukhi script documents is done manually. In this paper, we have presented a study of different features and classifiers combinations for text dependent writer identification model based on Gurmukhi text handwriting. This paper is divided into six sections. The introduction and related work have been presented in Sect. 1. Section 2 illustrates the Gurmukhi script and data collection phase. Section 3 portrays the feature extraction techniques. In Sect. 4, we have briefly discussed about classification techniques. Section 5 incorporates experimental results using these features and classifiers. At long last, conclusion and future works are exhibited in Sect. 6.

2 Gurmukhi Script and Data Set

Gurmukhi script is the script used for writing the *Punjabi* language. In Gurmukhi script, there are 35 basic character constants out of which the initial three are vowel bearers. In this work, we have considered all these 35 fundamental Gurmukhi characters. Twenty samples of each character written by each writer are taken. In this manner, we have collected 49,000 samples from 70 different writers. 70% data of 49,000 samples has been taken as training dataset and rest of data is considered as testing dataset.

3 Feature Extraction Techniques

The performance of writer identification system, basically, is dependent on the features that are being extracted. The extracted features ought to have the capacity to classify a writer in a unique way. In this work, we have explored various features like, zoning, diagonal, transition, intersection and open end points, centroid, horizontal peak extent, vertical peak extent, parabola curve fitting, and power curve fitting based features. These features are extracted by using a hierarchical technique presented by Kumar et al. [6]. Numerous scientists have been utilized zoning, diagonal, transitions and intersection point's, etc., based features for printed or handwritten character recognition work. Parabola curve fitting and power curve fitting based features are provided by Kumar et al. [7]. They have used these features for offline handwritten Gurmukhi character recognition. They have also proposed peak extent based feature extraction techniques for handwritten Gurmukhi character recognition [5]. In this work, we have considered these strategies for proposing a text dependent writer identification framework based on Gurmukhi text handwriting.

4 Classification Techniques

Classification is the last stage of the writer identification framework which is used to classify the writers based on the features extracted in the previous phase. In this stage, just a final decision is taken about the unknown writer of character, to which class it belongs to find the identity of a writer. In this paper, we have considered four different classification methods, namely, Naive Bayes, Decision Tree, Random Forest, and AdaBoostM1. The Naive Bayes [4] classifier is a basic method which has a very clear semantics representing a probabilistic knowledge. It assumes that in a given class, predicative attributes are conditionally independent. Attributes in decision tree are nodes and each leaf node is representing a class of the writer. Decision tree classifiers are used to classify various subsamples of dataset. Random forest is one of the most popular and powerful machine learning algorithms [1]. It is a type of ensemble machine learning algorithm. Random forest removed the over-fitting crisis of decision tree. The *meta* estimator that fits the number of decision tree classifiers for such purpose is called random forest. The random forest uses averaging to help in getting better predictive accuracy and control over-fitting. Random forest is unexcelled in accuracy among existing supervised learning algorithms for classification and runs efficiently on large data bases [12]. AdaBoostM1 is a machine learning *meta* algorithm and can be used in conjunction with many other types of learning algorithms to improve their performance. The output of the other learning algorithms is combined into a weighted sum that represents the final output of the boosted classifier. AdaBoostM1 is adaptive in the sense that subsequent weak learners are tweaked in favor of those instances misclassified by previous classifiers. In this work, we have considered AdaBoostM1 with random forest classifier for improve the proposed writer identification accuracy.

5 Experimental Results and Discussion

In this section, we have presented experimental results of various features and classifiers considered in this work. For experimental results, we have used a dataset of 49,000 samples collected from 70 different writers of isolated handwritten characters of Gurmukhi script. 70% data from 49,000 samples is taken as training dataset and rest of data is considered as testing dataset. Experimental results are derived using distinctive feature extraction methods and classification procedures. These outcomes are graphically portrayed in Fig. 1. We have accomplished most extreme identification accuracy of 81.75% with centroid features and AdaBoostM1 ensemble classification technique as depicted in Table 1. True Positive Rate (TPR) and False Positive Rate (FPR) of each writer for this case (Centroid features and AdaBoostM1 classification) are presented in Figs. 2 and 3.

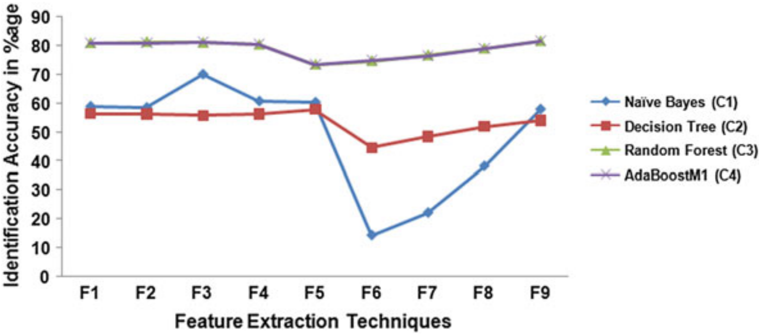


Fig. 1 Classifier-wise writer identification accuracy for different features

Table 1 Writer identification accuracy with different features and classifiers

	Classification Technique			
	Naive Bayes (C_1)	Decision Tree (C_2)	Random Forest (C_3)	AdaBoostM1 (C_4)
Zoning features (F_1)	58.93%	56.45%	81.04%	81.16%
Diagonal features (F_2)	58.54%	56.25%	81.46%	80.96%
Transition features (F_3)	70.10%	55.8%	81.19%	81.37%
Intersection and open end points based features (F_4)	60.75%	56.21%	80.49%	80.68%
Parabola curve fitting based features (F_5)	60.37%	57.84%	73.51%	73.57%
Power curve fitting based features (F_6)	14.40%	44.65%	74.78%	74.89%
Horizontally peak extent based features (F_7)	22.24%	48.45%	76.77%	76.59%
Vertically peak extent based features (F_8)	38.29%	51.88%	79.07%	79.21%
Centroid features (F_9)	57.97%	54.06%	81.70%	81.75%

6 Conclusion and Future Scope

In this paper, we have presented a study of different features and classification techniques for text-dependent writer identification system. Maximum identification accuracy of 81.75% has been accomplished with centroid features and AdaBoostM1 ensemble classifier. This accuracy may be improved either by increasing the span of training dataset or by using the various optimal feature selection techniques like PCA, Correlation Feature Set (CFS), etc. This work can be employed to other scripts like Devanagari, Bengali, and Tamil and so forth which are similar to the Gurmukhi script after building the training dataset of these scripts.

Fig. 2 True Positive Rate with Centroid Features and AdaBoostM1 Classifier

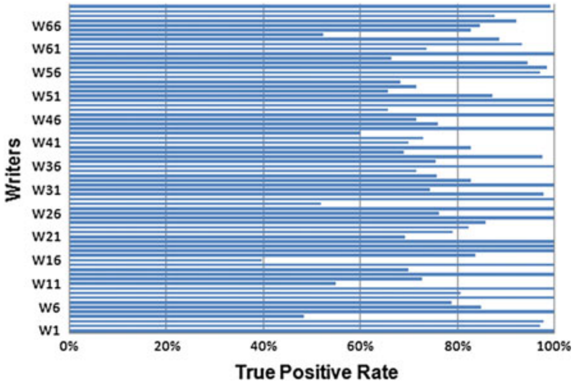
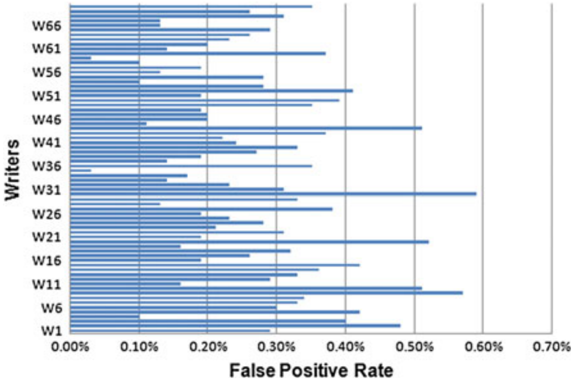


Fig. 3 False Positive Rate with Centroid Features and AdaBoostM1 Classifier



7 Declaration

In this manuscript, we have used 49,000 samples of Gurmukhi characters collected from 70 different writers. These all individuals who participated in this work have given their consent for publish this dataset in this manuscript.

References

1. Breiman L 2001 Random Forests, *Machine Learning*, 45(1):5–32.
2. Gazzah S and Amara N B 2008 Neural networks and support vector machines classifiers for writer identification using Arabic script, *The International Arab Journal of Information Technology*, 5(1): 92–101.
3. Ghiasi G and Safabakhsh R 2010 An efficient method for offline text independent writer identification, *In Proceedings of the 20th International Conference on Pattern Recognition*, 1245–1248.

4. John G H and Langley P 1995 Estimating Continuous Distributions in Bayesian Classifiers, *In Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, 338–345.
5. Kumar M, Sharma R K and Jindal M K 2013 A Novel Feature Extraction Technique for Offline Handwritten Gurmukhi Character Recognition, *IETE Journal of Research*, 59(6): 687–692.
6. Kumar M, Jindal M K and Sharma R K 2014a A Novel Hierarchical Techniques for Offline Handwritten Gurmukhi Character Recognition, *National Academy Science Letters*, 37(6): 567–572.
7. Kumar M, Sharma R K and Jindal M K 2014b Efficient Feature Extraction Techniques for Offline Handwritten Gurmukhi Character Recognition, *National Academy Science Letters*, 37(4):381–391.
8. Leclerc F, Plamondon R 1994 Automatic signature verification: the state of the art 1989–1993, *International Journal of Pattern Recognition and Artificial Intelligence,* 8(3):643–660.
9. Leeham G, Chachra S 2003 Writer identification using innovative binarised features of handwriting numerals, *In the Proceedings of the 7th International Conference on Document Analysis and Recognition* (ICDAR).
10. Maadeed S A 2012 Text-dependent writer identification for Arabic Handwriting, *Journal of Electrical and Computer Engineering*, 13: 1–8.
11. Plamondon R, Lorette G 1989 Automatic Signature Verification and Writer Identification The State of the Art, *Pattern Recognition*, 22(2):107–131.
12. Schlapbach A, Bunke H 2005 Writer identification using an HMM based hand writing recognition system: to normalize the input or not?, *In the Proceedings of 12th Conference of the International Graphonomics Society*, Salerno, Italy.
13. Schlapbach A, Bunke H 2007 A writer identification and verification system using HMM based recognizers, *Pattern Analysis and Application*, 10(1):33–43.
14. Zois E, Anastassopoulos V 2000 Morphological Waveform Coding for Writer Identification, *Pattern Recognition*, 33(3):385–398.