Applied Soft Computing 46 (2016) 980–990

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Robust soft-biometrics prediction from off-line handwriting analysis

Nesrine Bouadjenek*, Hassiba Nemmour, Youcef Chibani

LISIC. Lab, Faculty of Electronics and Computer Sciences, University of Sciences and Technology Houari Boumediene (USTHB), Algiers, Algeria

ARTICLE INFO

Article history:
Received 14 June 2015
Received in revised form 26 August 2015
Accepted 11 October 2015
Available online 28 November 2015

Keywords:
Handwriting recognition
Soft-biometrics
Fuzzy MIN-MAX combination
GLBP
SVM

ABSTRACT

Currently, writer's soft-biometrics prediction is gaining an important role in various domains related to forensics and anonymous writing identification. The purpose of this work is to develop a robust prediction of the writer's gender, age range and handedness. First, three prediction systems using SVM classifier and different features, that are pixel density, pixel distribution and gradient local binary patterns, are proposed. Since each system performs differently to the others, a combination method that aggregates a robust prediction from individual systems, is proposed. This combination uses Fuzzy MIN and MAX rules to combine membership degrees derived from predictor outputs according to their performances, which are modeled by Fuzzy measures. Experiments are conducted on two Arabic and English public handwriting datasets. The comparison of individual predictors with the state of the art highlights the relevance of proposed features. Besides, the proposed Fuzzy MIN-MAX combination comfortably outperforms individual systems and classical combination rules. Relatively to Sugeno's Fuzzy Integral, it has similar computational complexity while performing better in most cases.

1. Introduction

Handwriting recognition plays essential roles in various life domains such as mail sorting and bank checks verification. With the new technologies it is increasingly sought in more specific applications including information retrieval in historical documents and soft-biometrics prediction. Soft-biometrics is all what our senses perceive to differentiate us from each others, such as the age range, eye color, gender and ethnicity. These constitute key demographic attributes, which help to classify the human being into categories. During the last years, soft-biometrics traits were systematically predicted from face images [1,2]. Currently, there is a significant number of organizations that already employ handwriting analysis for personality profiling [3,4]. In fact, either for forensic identification of anonymous writing author, or the attribution of historical handwritten documents, soft-biometrics can be extremely useful. Furthermore, various studies tried to explain how the gender can control the human behavior. Specifically, the gender impact has been proved in Parkinson disease [5], motor learning [6], dichotic listening [7] as well as in crimes and violence [8]. Therefore, researchers in handwriting recognition were faced to a straightforward question, that is: Can the gender and other soft-biometrics influence the handwriting? In [9], authors investigated the relationship between sex hormones and the handwriting style. Their findings showed that prenatal sex hormones can affect the women handwriting. In some earlier psychological investigations, differences between men's and women's handwriting were examined [10,11]. Besides, in [12], experts were asked to predict the writer's gender from handwritten sentences. Experiments reported prediction accuracy about 68%. Also, in [13–15], age impact over the handwriting performance was investigated while in [16,17], authors tried to highlight the relationship between handedness and language dominance.

In handwritten document analysis field, automatic soft-biometrics prediction constitutes a new research subject. The literature reports only few studies, which addressed gender, handedness, age range and nationality prediction. The first work was developed in 2001 by Cha et al. [18]. Thereafter some other works were reported in [19–22].

A prediction system is composed of two main steps, that are feature generation and classification. In each step, efficient methods are required to achieve satisfactory performance. The key idea for developing a handwriting recognition system, is the choice of feature generation and classification schemes. Regarding the recognition step, a large number of classifiers that are based on different concepts such as singular value decomposition, principle component analysis, statistical modeling, as well as support vector methods are widely used for handwriting recognition [23]. In previous works on soft-biometrics prediction, various robust classifiers such as neural networks, SVM and Random Forests were employed while the feature generation was based on conventional direction, curvature and edge features. Note that SVM are considered as the best choice in most of recognition tasks where they commonly outperform other learning machines, namely, neural networks and HMM [24,25]. In fact, SVM are based on structural risk minimization, which answers two main problems of the statistical learning theory that are overfitting and controlling the classification complexity [26]. In addition, their training formulation is perfectly adequate to handle data with very large size without requiring dimensionality reduction. Furthermore, gender prediction results reported in [27] reveal that one SVM classifier could outperform the combination of multiple systems if they employ weak descriptors. Therefore, a straightforward way to achieve an efficient prediction is to associate robust data features to SVM.

This work is focused on the use of effective topological and gradient features, which are more suitable for handwriting characterization. Considered topological features are pixel density and pixel distribution, which gave satisfactory performance in handwritten signature verification [28]. As gradient feature the so called Gradient Local Binary Patterns (GLBP) is used. This descriptor was recently introduced for human detection in order to improve the histogram of oriented gradients. Three SVM predictors based respectively, on pixel density, pixel distribution and GLBP, are developed. Subsequently, a Fuzzy MIN-MAX combination algorithm

* Corresponding author. Tel.: +213 551544573.
E-mail addresses: nbouadjenek@usthb.dz (N. Bouadjenek), hnemmour@usthb.dz (H. Nemmour), ychibani@usthb.dz (Y. Chibani).

http://dx.doi.org/10.1016/j.asoc.2015.10.021
1568-4946/

is proposed to aggregate a robust prediction. Experimental analysis is carried out on public Arabic and English handwriting datasets.

The rest of this paper is as follows: Section 2 gives a brief description about the related work. Section 3 presents an overview of the proposed methods utilized to develop the prediction system. Section 4 states the problem of classifier combination in soft-biometrics. Section 5 introduces the proposed Fuzzy MIN-MAX combination algorithm. Section 6 details experimental results along with the experimental setup and computational complexity evaluation. Discussions regarding performance of the proposed combination algorithm and its comparison with the state of the art methodologies is placed in Section 7. The main conclusions are given in Section 8.

2. Related work

The first work on writer's soft-biometrics prediction was published by Cha et al. [18]. The aim was to classify the US population into various sub-categories such as white/male/age group 15–24 and white/female/age group 45–64. Experiments were conducted on CEDAR letter database, which contains 3000 handwritten document images written by 1000 subjects representative of the US population. A corpus of 200 samples was collected by considering six properties in defining categories that are: gender, handedness, age range, ethnicity, highest level of education and the place of schooling. Classical features such as pen pressure, writing movement, and stroke formation, were used with artificial neural networks. Experiments reveal a performance of 70.2% and 59.5% for gender and handedness prediction, respectively. Next, boosting techniques were employed to achieve higher performance where accuracies reached 77.5%, 86.6% and 74.4% for gender, age and handedness classification [29]. Thereafter, a research group on computer vision and artificial intelligence at Bern University developed the IAM handwriting dataset, which is developed for writer identification as well as gender and handedness prediction [30]. Authors utilized a set of 29 on-line and off-line features associated to SVM and Gaussian Mixture Models (GMM) [19,20]. On-line features cover several writing aspects such as the speed and acceleration, writing direction, normalized x and y coordinates, the vicinity curliness and the deviation of the straight line. Off-line features are based on conventional structural traits such as ascenders, descenders and the number of points above or below the corpus line. The handedness prediction using GMM classifier is achieved with an overall precision about 84.66%. The best gender prediction accuracy that is 67.57% was obtained by combining GMM trained separately over on-line and off-line features by using the average rule. In [27], similar gender prediction experiments were conducted by using more effective off-line descriptors, that are Local Binary Patterns (LBP) and Histogram of Oriented Gradients (HOG). SVM classifiers were used to perform the prediction task. HOG features provided the best prediction accuracy that is about 74%.

Furthermore, Al-Maadeed et al., [21] employed a K-Nearest Neighbors algorithm for handedness detection from off-line handwriting. A set of direction, curvature, tortuosity and edge-based features was used. The experimental dataset was collected at Qatar University by asking 1017 writers to reproduce two texts in both English and Arabic languages [31]. Then, the same features were used for gender, age range and nationality prediction in [22]. As a prediction method, Random Forest and Kernel Discriminant Analysis were used. Each classifier was trained using individual features and subsequently, various feature combinations were tested. The main conclusion that can be inferred from the experimental findings is that the feature combination that gives the best prediction accuracy is not the same for all soft-biometrics traits. For instance, an accuracy of 74% was collected for gender prediction by using probability density function of the chain code as a single feature, while for age range, the best accuracy reaches 62.5% by combining direction, curvature and tortuosity features. Also, weak prediction scores that are less than 50% were obtained for the nationality classification. Not long after that, using this dataset and another set of Arabic and French handwritten text, Siddiqi et al. [32] investigated gender classification using curvature, fractal and textural features. The classification was based on neural networks and SVM classifiers. Experiments showed that feature combination at the input of each classifier does not bring a significant improvement compared to individual features. Moreover, similar results are obtained by achieving either text-dependent or text-independent prediction. Unfortunately, in both works, datasets are not publically available to perform comparison.

The inspection of all previous works reveals that predicting writer's soft-biometrics is a very complicated task, since the classification scores are commonly around 70%. Such results let soft-biometrics prediction an open research area where a lot of work could be done in both feature extraction and classification.

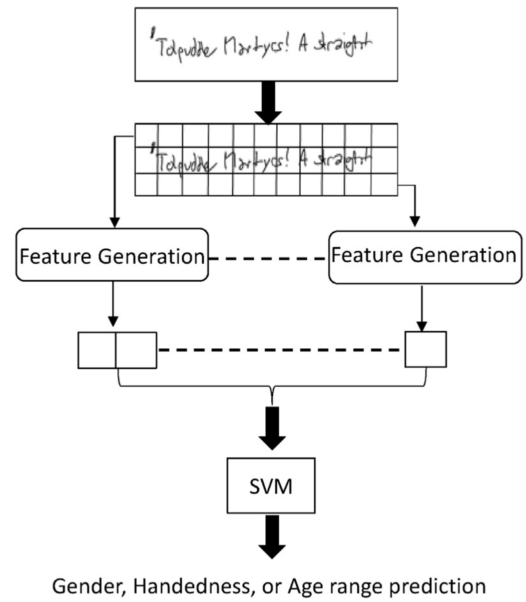

Fig. 1. Proposed system for soft-biometrics classification.

3. Proposed systems for soft-biometrics classification

Soft-biometrics classification systems are designed to automatically classify writers into specific categories such as "man" or "woman" in the case of gender prediction, "left hand" or "right hand" for handedness prediction and various age ranges, in the case of age prediction. Similarly to all handwriting recognition systems, two main steps, that are feature generation and classification, are required. As shown in Fig. 1, features are locally extracted by applying grid over text images. Then, the feature vector of the full text image is obtained by concatenating all cells features.

3.1. Feature sets

3.1.1. Topological features

Two grid-based features, namely, pixel density and pixel distribution are used to highlight topological properties of handwritten data. The density is what we call apparent pressure, since it describes the width of the strokes. This feature is obtained by considering the ratio between the number of pixels that belong to the text and the cell's size. As reported in [28,27], within a given cell, the distribution is based on four measures that are: The heights of the left and the right parts of the stroke (designated by A and C in Fig. 2), and the widths of the upper and lower parts of the stroke (designated by B and D in Fig. 2).