Expert Systems With Applications 85 (2017) 158–168

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Improving handwriting based gender classification using ensemble classifiers

Mahreen Ahmed[a], Asma Ghulam Rasool[a], Hammad Afzal[a], Imran Siddiqi[b,*]

[a] National University of Sciences and Technology, Islamabad, Pakistan
[b] Center of Computer Vision and Pattern Recognition, Bahria University, E8, Islamabad, Pakistan

ARTICLE INFO

Article history:
Received 30 July 2016
Revised 8 April 2017
Accepted 13 May 2017
Available online 15 May 2017

Keywords:
Handwritten documents
Gender classification
Classifier combination
Textural features

ABSTRACT

This paper presents a system to predict gender of individuals from offline handwriting samples. The technique relies on extracting a set of textural features from handwriting samples of male and female writers and training multiple classifiers to learn to discriminate between the two gender classes. The features include local binary patterns (LBP), histogram of oriented gradients (HOG), statistics computed from gray-level co-occurrence matrices (GLCM) and features extracted through segmentation-based fractal texture analysis (SFTA). For classification, we employ artificial neural networks (ANN), support vector machine (SVM), nearest neighbor classifier (NN), decision trees (DT) and random forests (RF). Classifiers are then combined using bagging, voting and stacking techniques to enhance the overall system performance. The realized classification rates are significantly better than those of the state-of-the-art systems on this problem validating the ideas put forward in this study.

1. Introduction

Handwriting analysis has been vastly used to characterize writer's identity since pre-computers era. One of the commonly used practices is signature verification in banking system wherein signatures provided by any particular writer are used as his/her identity. This is plausible because, despite the fact that most of the people learn writing from given printed shapes in books, each writer develops his/her own style with time (Koppenhaver, 2007). This has resulted in interest of neurologists, forensic experts and document examiners in hand writing analysis. Handwriting represents a complex fine motor skill (Caligiuri & Mohammed, 2012; Feder & Majnemer, 2007) involving cognitive, psychomotor and biophysical processes (Van Galen, 1991). Correlation has been shown to exist between handwriting and a number of neurological disorders like autism (Fuentes, Mostofsky, & Bastian, 2009; Kushki, Chau, & Anagnostou, 2011) Parkinson (Teulings & Stelmach, 1991) and Alzheimer (Schröter et al., 2003). Likewise, effects of psychotropic medications (Caligiuri & Mohammed, 2012) and aging (Rosenblum, Engel-Yeger, & Fogel, 2013; Walton, 1997) on handwriting have also been studied. Among different attributes, gender has shown to be significantly affecting the writing style of individuals (Beech & Mackintosh, 2005; Hamid & Loewenthal, 1996; Hartley, 1991; Hayes, 1996). The differences in writing styles (Koppenhaver, 2007) are generally attributed to different motor coordination (Dorfberger, Adi-Japha, & Karni, 2009; Hartley, 1991) and hormones (Hayes, 1996) in males and females.

Typical tasks based on analysis of handwriting include handwriting recognition, writer identification/verification and signature verification etc. These and other similar tasks offer useful applications including forensic, biometric and historical document analysis, criminal investigation, smart meeting systems, classification of ancient manuscripts and personalized handwriting recognition systems. Another interesting aspect is the prediction of gender and other demographic attributes of writers from handwriting. Certain writing traits have been identified by researchers which can be exploited to distinguish writings of male and female writers.

Schomaker (2008) attributes the differences in handwritings of individuals to biological and memetic attributes. The biological factors include structure of hand, handedness (left or right) (Francks et al., 2003), strength of muscles (Gulcher et al., 1997) and the stability of motor control (Van Galen, Van Doorn, & Schomaker, 1990). The memetic or cultural factors refer to the acquired styles of pen grip and drawing character shapes (allographs) which are learned through education or observation. While the writing process is a function of these genetic and memetic factors, gender based differences in handwriting are mostly linked to the motor control differences in males and females (Dorfberger, Adi-Japha, & Karni, 2009). Learning rate of motor skills is also known to vary in males and fe-

* Corresponding author.
E-mail addresses: mahreenmcs@gmail.com (M. Ahmed), asma.ghulamrasul@gmail.com (A.G. Rasool), hammad.afzal@mcs.edu.pk (H. Afzal), imran.siddiqi@bahria.edu.pk (I. Siddiqi).

http://dx.doi.org/10.1016/j.eswa.2017.05.033
0957-4174/

males (Buitrago, Ringer, Schulz, Dichgans, & Luft, 2004). These gender based differences in the central nervous system motor functioning are known to lead to differences in male and female writings (Cohen, 1997; Weintraub, Drory-Asayag, Dekel, Jokobovits, & Parush, 2007).

Recently with vast adoption of computers in every aspect of life, handwriting analysis is also being automated by using pattern recognition and image analysis algorithms. Most of such automatic techniques have been derived or are inspired from the ones reported through manual examination of handwriting (Goodenough, 1945; Hartley, 1991; Hayes, 1996). These techniques employ computationally measurable features such as slant, word spacing, pen pressure, gradient information etc. Computational features are those which contain known software and hardware methods for extraction. Computational features can further be divided into two categories: micro (character) and macro (document level) features. A number of techniques have been developed where different computational features are extracted from scanned images of handwriting employed with various classification algorithms to determine the gender. Some of the examples are Fourier descriptors, tangent orientation and curvature information etc. used along with classification techniques such as SVM, ANN (Al Maadeed & Hassaine, 2014; Liwicki, Schlapbach, & Bunke, 2011; Siddiqi, Djeddi, Raza, & Souici-meslati, 2014; Sokic, Salihbegovic, & Ahic-Djokic, 2012).

Major proportion of the research carried out on this subject focuses on adapting the features used for writer identification and handwriting recognition for gender classification. The standard features reported in the literature are extracted from writing samples of male and female writers and a traditional classifier like ANN or SVM is used to learn to discriminate between the two classes. This paper presents an improved method for gender prediction from offline handwriting images where the major focus of study is on improving the classification part. A set of descriptors mainly capturing the textural information in handwriting is exploited to characterize male and female writings. These features are then fed to different classifiers to recognize gender from handwriting. We explored various combinations of ensemble techniques such as bagging, voting and stacking to combine the individual classifiers. The proposed system evaluated on the standard QUWI database realizes high classification rates and demonstrates the effectiveness of the proposed ensemble technique over traditional classification. The results are also compared with a number of state-of-the-art techniques on this problem highlighting significant improvements in classification rates under similar experimental settings.

This paper is organized as follows. In the next section, we discuss the major recent developments reported in the literature for handwriting based gender classification. In Section 3, we introduce the proposed methodology including the feature extraction technique and the classifier ensemble employed. Section 4 presents a description of the database, the experimental setup, realized results and a discussion on the system performance. The last section concludes the paper with a discussion on potential research directions on this problem.

2. Related work

This section presents an overview of the research on prediction of gender using handwriting analysis ranging from pre-computers era using conventional features to the most recent automatic analysis using computational features. We have primarily focused on techniques that use offline data considering only those attributes of handwriting that are available in offline text. This is contrary to online features that are captured in real time, for instance, pen pressure, number and order of strokes and writing speed etc. Studies that involve demographic features other than gender such as age, handedness and race are also briefly discussed due to relevance of the problem.

Among one of the earliest works reported for gender prediction (Goodenough, 1945), characteristics of male and female writings are discussed from the perspective of psychologists and document examiners. Hartley (1991) identified some of the (conventional) features to discriminate among handwritings of males and females. In another study, Hayes (1996) studied the role of hormones in determining the handwriting style. Hamid and Loewenthal (1996) carried out a study to predict gender from handwriting for English and Urdu text. The authors created a dataset comprising writings of 30 (16 female and 14 male) writers for training and 25 (13 female and 12 male) writers for testing. Classification was carried out by human experts and an average classification rate of around 68% is reported in the study.

Among computerized analysis of handwriting, Cha and Srihari (2001) proposed a system for demographic classification of US population according to race, gender and age group (for example, 'white/male/age group 15–24') based on handwritten images. An overall classification rate of 70.2% is reported in the study. In a later study, Bandi and Srihari (2005) carried out age, gender and handedness prediction from CEDAR letters (800 samples for training and 400 for testing). The classification performance using boosting techniques with 10 NN realized best results where the classification performance peaked at 77.5%, 86.6% and 74.4% for gender, age and handedness classification, respectively.

In a relatively recent study, Liwicki, Schlapbach, and Bunke (2011) carried out classification of gender and handedness from the online handwritten text. The authors employed a set of online and off-line features and performed classification using Support Vector Machine and Gaussian Mixture Models. The technique was evaluated on 200 writers of the IAM-OnDB database with 8 samples per writer. The authors report a gender classification rate of around 67% and handedness classification rate of 85%. Sokic, Salihbegovic, and Ahic-Djokic (2012) employ a set of descriptors including tangent and curvature information and Fourier descriptors to discriminate between male and female writings. The authors evaluated the technique on a custom developed database and demonstrated the values of different features computed for the same word written by male and female writers.

Siddiqi, Djeddi, Raza, and Souici-meslati (2014) exploit writing attributes like slant, curvature, texture and legibility to differentiate between male and female writings. These attributes are captured by computing a set of global and local features including histograms of chain codes, fractal dimensions and local binary patterns. Classification is carried out using Support Vector Machine (SVM) and Artificial Neural Networks (ANN). The proposed system is evaluated on two different databases, QUWI and MSHD. Experiments are carried out through a number of interesting evaluation scenarios including script dependent and script independent, text dependent and text independent and cross database evaluations. Classification rates of 68–74% are realized in different experimental scenarios.

Al Maadeed and Hassaine (2014) investigated a set of geometric features to characterize the gender of writer. Classification is carried out using random forests and kernel discriminant analysis and experiments are carried out in text dependent and text independent modes on the QUWI database. In text dependent mode, the authors report classification rates of 74.05%, 55.76% and 53.66% for gender, age and nationality prediction, respectively. Likewise, for text independent experiments, classification rates of 73.59%, 60.62% and 47.98% are reported for gender, age range and nationality, respectively. In a later study (Al-Maadeed, Ferjani, Elloumi, & Jaoua, 2016), the authors employ dimensionality reduction based on fuzzy conceptual reduction using Lukasiewicz implication and demonstrate its effectiveness on handedness detection from offline

handwriting. The primary focus of the study is the reduction of the feature vector dimensionality and handedness detection is used as a case study. The proposed technique reduces the feature vector dimension by around 31% maintaining high classification rates.

Bouadjenek, Nemmour, and Chibani (2014) study the effectiveness of local descriptors including Histogram of Oriented Gradients (HOG), Local Binary Patterns (LBP) and pixel density in small grids using SVM as classifier to predict gender from offline images of handwriting. The system was evaluated on 200 writers of the IAM-OnDB database and a performance comparison of different features demonstrated that HOG features outperform other features realizing a classification rate of 74%. In a later study (Bouadjenek, Nemmour, & Chibani, 2015b) by same authors, the HOG descriptor was employed to predict age, gender and handedness on writing images in the QUWI and KHATT databases. HOG was also combined with a variant of LBP feature, the gradient local binary patterns (GLBP) to perform age, gender and handedness classification on IAM and KHATT databases (Bouadjenek, Nemmour, & Chibani, 2015a).

Summarizing the discussion on handwriting based gender classification, it can be observed that in many cases, features employed for identification of writers and handwriting recognition can be adapted and tailored for characterizing gender from handwriting images. Another interesting aspect could be to investigate the performance of features employed for classification of writing styles (Brink, Smit, Bulacu, & Schomaker, 2012; Joutel, Eglin, Bres, & Emptoz, 2007; Siddiqi, Cloppet, & Vincent, 2009; Yosef, Kedem, Dinstein, Beit-Arie, & Engel, 2004) in contemporary as well as ancient documents, for gender classification problem. In a number of recent studies, high level computer vision descriptors like Scale Invariant Feature Transform (SIFT) (Wu, Tang, & Bu, 2014; Xiong, Wen, Wang, & Lu, 2015) and bag of visual words models (Fiel & Sablatnig, 2013; Gordo, Fornés, & Valveny, 2013) have been adapted for writer identification problem. It would be interesting to study how these features can be tailored for the more general gender classification task.

After having discussed the notable recent contributions to handwriting based writer demographics classification, we present the proposed technique in the next section.

3. Proposed methodology

This section presents in detail the proposed technique to characterize gender from offline images of handwriting. We first discuss the feature extraction where we present the set of textural descriptors employed in our study. Later, we detail the individual classifiers as well as the different ensemble techniques including voting, bagging and stacking through which the individual classifiers are combined.

3.1. Feature extraction

Feature extraction aims to find a characteristic representation of handwriting images that allows to discriminate between male and female writers. We exploit the textural information of handwriting to differentiate between the two gender classes. The textural information is captured using a set of features including Segmentation-based Fractal Texture Analysis (SFTA), local binary patterns (LBP), histogram of oriented gradients (HoG) and features extracted from Gray-level co-occurrence matrices (GLCM). It should be noted that these features have been employed in a number of recognition problems. The objective of this study is not to propose novel features but to employ state-of-the-art features and enhance the classification rates using the ensemble classifiers. For completeness, a brief description of these features is presented in the following.

3.1.1. Segmentation-based fractal texture analysis

Fractal analysis is an effective way to characterize the texture in handwriting. A major issue, however, is the high computational cost of most of the fractal analysis techniques. In a relatively recent study, Costa, Humpire-Mamani, and Traina (2012) presented an efficient technique to extract fractal based textural information, the SFTA. The original image is used to produce multiple binary images using a Two-Threshold Binary Decomposition (TTBD) algorithm with two thresholds (upper and lower).

$$I_b(x,y) = \begin{cases} 1, & \text{if } t_l < I(x,y) \le t_u. \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

Where, t_l and t_u represent lower and upper threshold values respectively. Given T to be a set of threshold values obtained by selecting equally spaced gray levels and n_l to be the maximum gray value in the image, the image is binarized using all pairs of successive thresholds from $T \cup \{n_l\}$ and all pairs of thresholds $\{t, n_l\}$, $t \in T$ (Costa, Humpire-Mamani, & Traina, 2012). Consequently, the number of binary images is twice the number of thresholds in T. The SFTA feature vector is calculated from each of the binary images and comprises the size (pixel count), mean gray level value and fractal dimension. The fractal dimension is calculated using the Hausdorff's dimension based on box counting method (Traina Jr, Traina, Wu, & Faloutsos, 2010). The total dimension of the SFTA feature vector is the number of binary images produced by TTBD multiplied by three. In our implementation, we use 4 threshold values producing 8 binary images and a 24 dimensional feature vector.

3.1.2. Local binary patterns

Local binary patterns, originally proposed in (Ojala, Pietikainen, & Harwood, 1994) for texture classification, have been successfully applied to a number of classification problems (Bertolini, Oliveira, Justino, & Sabourin, 2013; Siddiqi, Djeddi, Raza, & Souici-meslati, 2014; Wang, Han, & Yan, 2009). The original LBP operator considers the 3×3 neighborhood of each pixel and using the central value as threshold, binarizes the 8 neighboring pixels producing a string of 0s and 1s, a binary number. Each pixel is then labeled with the resulting number and a histogram of these values is used as a texture descriptor. Later, the authors proposed a number of extensions in the basic LBP operator (Ojala, Pietikainen, & Maenpaa, 2002). The notable of these include taking into account neighborhoods of different sizes and the concept of uniform (less than 2 transitions between 0s and 1s) and non-uniform (more than 2 transitions) binary patterns. Generally, there is a separate bin for each uniform pattern and all the non-uniform patterns are counted in a single bin.

In our implementation, we compute the (16, 2) LBP, i.e., 16 neighboring pixels at a distance of 2 pixels from the central pixel. For 16 neighboring points, there is a total of 242 uniform patterns. The LBP histogram, therefore, has 243 bins, 242 for the uniform patterns and 1 for all the non-uniform patterns.

3.1.3. Histogram of Oriented Gradients

Histogram of Oriented Gradient (HoG) feature descriptor, vastly used for object detection, exploits gradient orientation in localized parts of images. First introduced by Dalal and Triggs (2005), these features are suitable to detect local changes in position and appearance and have been successfully applied to a number of recognition tasks (Déniz, Bueno, Salido, & De la Torre, 2011; Kobayashi, Hidaka, & Kurita, 2008; Minetto, Thome, Cord, Leite, & Stolfi, 2013; Rybski, Huber, Morris, & Hoffman, 2010; Sharma, Pal, & Blumenstein, 2014; Wang, Han, & Yan, 2009). The key idea of HoG is to characterize an object by the distribution of edge (gradient) directions. In a number of recent studies, HoG features have shown ef-

Table 1
Summary of GLCM based features.

SNo.	Feature	Computational details
1.	Contrast	$\sum_{i,j=0}^{N-1} P_{i,j} = (i,j)^2$
2.	Correlation	$\sum_{i,j=0}^{N-1} P_{i,j} \left[\frac{(i-\mu_i)(j-\mu_j)}{(\sqrt{(\sigma_i^2)(\sigma_j^2)})} \right]$
3.	Homogenity	$\sum_{i,j=0}^{N-1} P_{i,j} = 0 \frac{P_{i,j}}{i+(i-j)^2}$
4.	Entropy	$\sum_{i,j=0}^{N-1} P_{i,j}(-lnP_{i,j})$
5.	Energy	$\sum_{i=0}^{N-1} \sum_{j=0}^{N-1} [P(i,j)]^2$

Table 2
Summary of features.

SNo.	Feature	Dimension
1.	SFTA	24
2.	LBP	243
3.	HoG	81
4.	GLCM Statistics	20

fective performance for analysis of handwritten texts (Bouadjenek, Nemmour, & Chibani, 2015a; 2015b; Ebrahimzadeh & Jampour, 2014). We, therefore, employ HoG descriptor as one of the features characterizing the gender of writer.

From the view point of implementation, the handwritten image is divided into cells. For each pixel within a cell, the gradient vector (magnitude and direction) is computed. A histogram of orientations is then computed for each cell and the descriptor for the complete image is calculated by concatenating the histograms of all the cells.

3.1.4. Gray-level co-occurrence matrices

GLCMs represent an effective textural measure. GLCMs consider the relationship among two neighboring pixels and determine how frequently different combinations of gray levels co-occur for a given direction and distance. Formally, the GLCM $P(i, j)$ is defined by first specifying a displacement vector $d = (dx, dy)$ and counting all pairs of pixels separated by d having gray levels i and j. The size of GLCM matrix is the same as the number of gray levels in the image. In our implementation, we consider binary images of handwriting and compute the GLCMs using four displacement vectors (offsets). These offsets include (0,1), (1,−1), (0,−1) and (−1,−1) and correspond to four directions 0°, 45°, 90°, 135°.

Texture is generally characterized by computing statistics from GLCMs. In our study, we employ the well-known statistics of contrast, correlation, homogeneity, entropy and energy of each GLCM and use them as features to characterize each block. These features are summarized in Table 1. These five statistics are computed for each of the four GLCMs (0°, 45°, 90°, 135°) giving a 20 dimensional feature vector.

Table 2 provides a summary of the texture features employed in our study along with the dimensionality of each.

3.2. Classification

Classification is carried out using individual learners as well as the ensemble learners as discussed in the following.

3.2.1. Individual classifiers

For comparison purposes, we applied different individual learners including Decision Trees (DT), Support Vector Machine (SVM), K Nearest Neighbor (KNN), Random Forests (RF) and Artificial Neural Network (ANN) on the four features (LBP, HOG, GLCM, and SFTA) individually and on the combination of these features (LBP+HOG+GLCM+SFTA).

3.2.2. Ensemble classifier or metaLearner

An ensemble classifier is a set of classifiers that classify the set of samples by taking vote on the predictions of its constituents. Over the years ensemble classifiers have been used to obtain high accurate classifiers by combining the less accurate ones (Dietterich, 2000). While combining multiple classifiers, there are generally two types of approaches: classifier selection and classifier fusion (Ho, Hull, & Srihari, 1994). In classifier selection, a single classifier is selected that gives the best accurate results for a given sample. In classifier fusion, different classifiers are applied in parallel to get the group's consent on the final decision. Thus each learner has some part in this decision as opposed to the classifier selection (Woods, Bowyer, & Kegelmeyer Jr, 1996).

In our study, we applied ensemble learners including Bagging (BAG), Majority Voting (MV) and Stacked Generalization (SG) on the four features individually and on the combination of these features.

Bagging. Bagging, also known as Bootstrap Aggregating, is a technique that is used to enhance the performance of the base classifier. Bagging makes bootstrap samples B_i (where $i = 1, \ldots, n$) of the dataset D and applies the learning algorithm to each sample, averaging the classification results in the end (Breiman, 1996). In bootstrap, samples are selected randomly from the data set with replacement. We get n bootstrapped sample sets when we repeat the process n times. As the samples are selected randomly there is a probability of $1/n$ times of a sample being selected each time. Each sample may be repeated a number of times or not appear at all in a bootstrapped sample set.

Bagging is a voting method that uses bootstrap to generate sample sets and for any given instance, the class chosen by most of the classifiers is voted as the ensemble decision. Bagging trains the base learner using an unstable learning approach. A learning algorithm is said to be unstable if it shows a significant improvement in accuracy with slight changes in the trained dataset (Alpaydin, 2014). For unstable learning algorithms like Decision Trees (DT) and Artificial Neural Networks (ANN), small changes in training data will give different results while stable learners like K- Nearest Neighbor (KNN) and Linear Regression (LR) show no or little changes (Zhang & Tsai, 2005). Bagging is particularly interesting when the data is of an inadequate size. Usually a large portion of the samples are drawn into each bootstrapped subset. This causes distinct training subsets to overlap significantly, with many of the same samples appearing in most of the bootstrapped subsets. Some of the samples appear multiple times in a subset. To ensure diversity, the unstable learner is used so that different decision boundaries can be obtained for small disruptions in different training datasets. For this purpose, ANNs and DTs are good candidates. The instability of such unstable learners can be controlled by the selection of the free parameters. That is why bagging is usually applied to DT algorithms. Bagging methods needs a preprocessor that takes the bootstrap replicates of the set to the unstable learner and a postprocessor that aggregates the votes (Liatsis, 2002). Bagging is one of the most instinctive and simplest to implement and has a surprisingly good performance. Fig. 1 shows the methodology of the bagging classifier.

In our experiments with bagging, we performed bagging on individual and meta learners. Five experiments were performed including:

- Bagging with SVM (BAG-SVM)
- Bagging with KNN (BAG-KNN)
- Bagging with ANN (BAG-ANN)
- Bagging with DT (BAG-DT)
- Bagging with Random Forest (BAG-RF)

More details can be found in the next section.

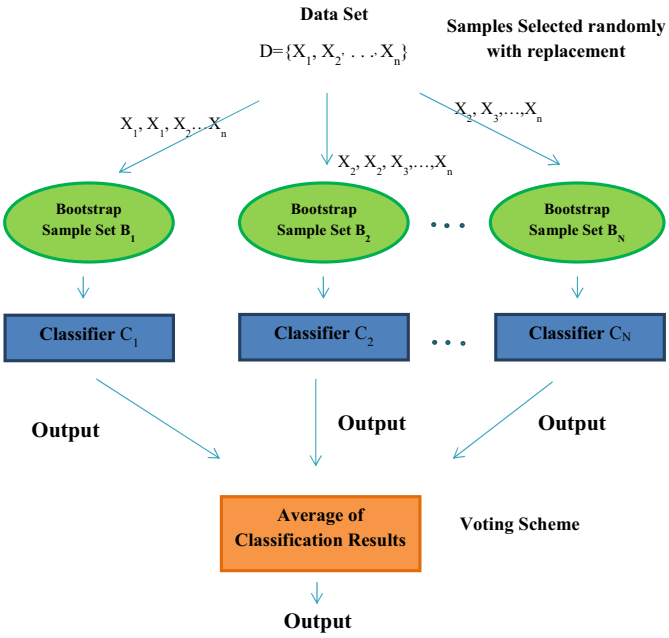

Fig. 1. General idea of bagging.

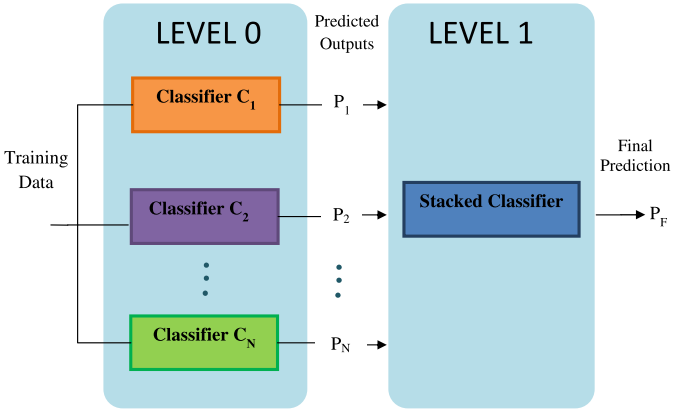

Fig. 2. Stacked generalization.

Stacked generalization or stacking. Stacking, an ensemble classification algorithm, is the combination of different heterogeneous learning algorithms that are applied on the same data set. It is a meta-model that has base models referred as Level-0 and the meta-model referred as Level-1 model. Level-1 model combines the set of outputs of the Level-0 models and rectifies their mistakes to improve the classification results (Sammut & Webb, 2011). Stacking creates a meta-dataset, containing the tuples of the original dataset using the predictions made by the classifiers as input attributes. The target attribute is the same for both the original and stacked dataset (Acton, 2013). The meta-model combines the different predictions into a single prediction as seen in Fig. 2.

Stacked generalization or stacking was proposed by David H. Wolpert and is described as a method for reducing the error rate of one or more classifiers or generalizers C_i (where $i = 1, \ldots, n$) like decision trees or neural networks. Stacked generalization guesses the bias of a generalizer C_i within the data set D provided for training and the outputs of these generalizers serve as input along with the learning set to get the correct guess. When used with more than one classifier, it is a refined version of cross validation. Generally stacking has two levels. The space occupied by the original learning set D is the level 0 space and the generalizers C_i used to train the learning set in this space are the level 0 generalizers. The first step in stacking is to partition the learning set into training $D1$ and test set $D2$. The outputs for $D1$ from the generalizers C_i make the input for $D2$ and are fed to a generalizer in the level 1 space. Thus stacking takes a question in the level 0 space and passes the answer to level 1 space to generate a level 1 question which is answered by the level 1 learning set. Wolpert (1992) discusses the fact that the selection of the generalizers in the level 0 space is not defined. There is no specific list of generalizers that should be used for stacking. In another study, Ting and Witten (1997) concluded that stacked generalization works best with the combination of three heterogeneous base learners and can achieve better accuracy than majority voting. Džeroski and Ženko (2004) discuss the influence of the number of base classifiers and conclude that three base classifiers perform the same as seven base classifiers. There is not much difference in the results and in fact some stacking models perform much better with three base classifiers. One of the advantages of stacking is the free choice of base classifiers. One can use a single classifier or multiple classifiers. Researchers have been investigating the best methods for constructing the ensemble classifiers with stacking.

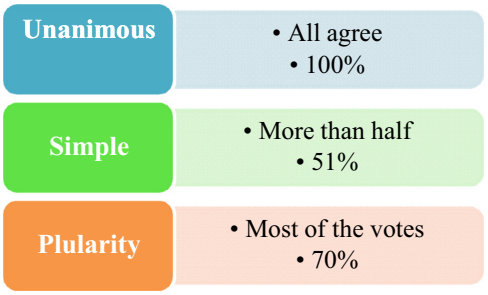

Fig. 3. Types of majority voting.

Stacking with two or three level-0 base learners was employed in our experiments. Six experiments were performed with stacking.

- Stacking with SVM, KNN and DT (SG-SVM-KNN-DT): SVM, KNN were used as base learners at level-0 and DT was used at level-1.
- Stacking with ANN, SVM and DT (SG-ANN-SVM-DT): ANN, SVM were used as base learners at level-0 and DT was used at level-1.
- Stacking with ANN, SVM and KNN (SG-ANN-SVM-KNN): ANN, SVM were used as base learners at level-0 and KNN was used at level-1.
- Stacking with SVM, DT and KNN (SG-SVM-DT-KNN): SVM, DT were used as base learners at level-0 and KNN was used at level-1.
- Stacking with SVM, KNN, DT and KNN (SG-SVM-KNN-DT-KNN): SVM, KNN, DT were used as base learners at level-0 and KNN was used at level-1.
- Stacking with SVM, KNN, ANN and DT (SG-SVM-KNN-ANN-DT): SVM, KNN and ANN were used as base learners at level-0 and DT was used at level-1.

Majority voting. Majority voting can be classified as (i) unanimous (ii) simple and (iii) plurality voting. Unanimous means that all classifiers agree, simple voting signifies that at least more than half agree, plurality means the highest votes received as seen in Fig. 3. Plurality voting is the most common technique used when it comes to majority voting (Polikar, Zhang, & Ma, 2012). Plurality voting and simple majority voting are identical when more than half of the learners agree on the same prediction. This is similar to the voting mechanism used in audience polling.

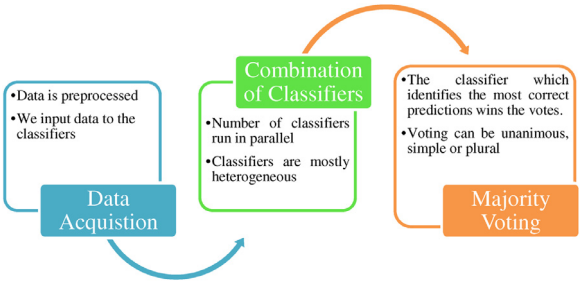

Fig. 4. Majority voting.

Majority voting is an ensemble technique in which we use N base classifiers as explained in Fig. 4. The result is mostly the common output generated by each base classifier. Majority voting will only get the answer wrong if more than half of the base classifiers give the wrong answer. Majority voting returns the correct label when there are odd numbers of base classifiers, they are independent of each other and more than half of these classifiers agree on the same answer. Suppose there are N classifiers and each has the probability p to predict the correct class. The voting method makes the correct decision when $N/2+1$ of these classifiers gives the correct label (Marsland, 2014). This combining scheme is mostly used for comparing the newly proposed methods. We can give the more accurate learner more power if the separate classifiers do not give the same accuracy. This is termed as weighted majority voting. Weighted Majority Voting gives better results than the single learner and the simple majority voting (Kuncheva, 2004).

Majority voting with three or four base learners was employed in our study. Five experiments were performed with majority voting as listed in the following.

- Majority Voting with KNN, SVM and ANN (MV-KNN-SVM-ANN).
- Majority Voting with KNN, SVM and DT (MV-KNN-SVM-DT).
- Majority Voting with SVM, DT and ANN (MV-SVM-DT-ANN).
- Majority Voting with KNN, DT and ANN (MV-KNN-DT-ANN).
- Majority Voting with SVM, ANN, DT and KNN (MV-SVM-ANN-DT-KNN).

4. Experiments and results

The experimental evaluation of the system was carried out on a subset of the QUWI database (Al Máadeed, Ayouby, Hassaine, & Aljaam, 2012) which comprises collections of handwritten samples contributed by a total of 1017 writers. Each writer contributed 4 samples, 2 in English and 2 in Arabic. The first sample of each writer comprises an arbitrary text in Arabic while the second sample contains a pre-defined Arabic text copied by all writers. Likewise, the third and fourth sample of each writer contains an arbitrary and fixed text in English respectively. This allows usage of this database in different interesting evaluation scenarios (Siddiqi, Djeddi, Raza, & Souici-meslati, 2014) like text-dependent vs. text-independent and scrip-dependent vs. script-independent experiments etc. Recently, the International handwriting-based gender classification competition held in conjunction with ICDAR2015 (Djeddi et al., 2015) also employed the QUWI database for evaluations. Writing samples of male and female writers in the database are illustrated in Fig. 5.

In order to allow meaningful comparison of the performance of our system with recently developed techniques on this subject, we carry out a series of evaluations using the same experimental setup as in the ICDAR2015 competition. Four scenarios were considered in the competition as outlined in the following.

- Scenario A: Gender classification on Arabic writings.
- Scenario B: Gender classification on English writings.
- Scenario C: Gender classification using Arabic samples in training and English samples in test.
- Scenario D: Gender classification using English samples in training and Arabic samples in test.

For each of the above scenarios, 300 writing samples were used as training, 100 as validation and 100 as test set. It should also be noted that no writers were common in the training and test sets. Scenarios A and B correspond to script dependent experiments where writing samples in training and test sets are in the same script (Arabic or English). Scenarios C and D, on the other hand, represent script independent evaluations where the training and test samples are in different scripts and hence are more challenging.

4.1. Performance of individual classifiers

We first present the performance of individual classifiers on each of the four scenarios. The classification rates are presented for each of the individual features as well as their combination. Table 3 summarizes the classification rates of individual classifiers for Scenarios A and B while Table 4 presents these rates for Scenarios C and D. The bold values in these (and subsequent) tables represent, for each feature (and feature combination), the highest realized classification rate.

It can be seen from Table 3 that for script dependent experiments, in general, the classification rates are higher on English writings as opposed to Arabic. This may be attributed to the highly cursive nature of Arabic handwriting making it challenging for the extracted features to characterize the gender of a given sample. Comparing the performance of individual features, the texture measures LBP and GLCM statistics outperform HoG and SFTA based features. This observation is very much natural as LBP and GLCM are known to be more effective textural descriptors as compared to HoG and SFTA and the proposed idea relies mainly on texture as the discriminating attribute between male and female writings. Comparing the performance of different classifiers, in general, ANN and SVM report higher classification rates and the trend is more or less consistent for all the individual features as well as their combination. In scenario A, the best performing classifier is ANN for individual and combined features. The next best performing classifier is SVM while DT, KNN and RF report acceptable performances as well. In Scenario B, ANN realizes the best results for individual and combined features as well. In Scenario C, ANN reports the highest classification rate of 79% for all features while for Scenario D, 80% classification rate using GLCM is realized. For all scenarios, the individual classifiers ANN and SVM report high classification rates on GLCM, LBP and combined features.

Comparing the system performance of script-dependent (Table 3) and script-independent evaluations (Table 4), it can be seen that the classification rates for experimental scenarios C and D are comparable to those for scenarios A and B. These classification rates validate the idea that individuals do share some common writing characteristics across different scripts. Having training and test samples in different scripts (Arabic and English) and still realizing comparable classification rates is indeed very promising.

4.2. Performance of ensemble classifiers

After having presented and discussed the performance of the individual classifiers, we now present the performance of the ensemble classifiers. Evaluations were carried out for each of the four experimental scenarios (A, B, C and D) using each of the combinations presented in Section 3.2.2. The classification rates of different classifier combinations for each of the individual features as

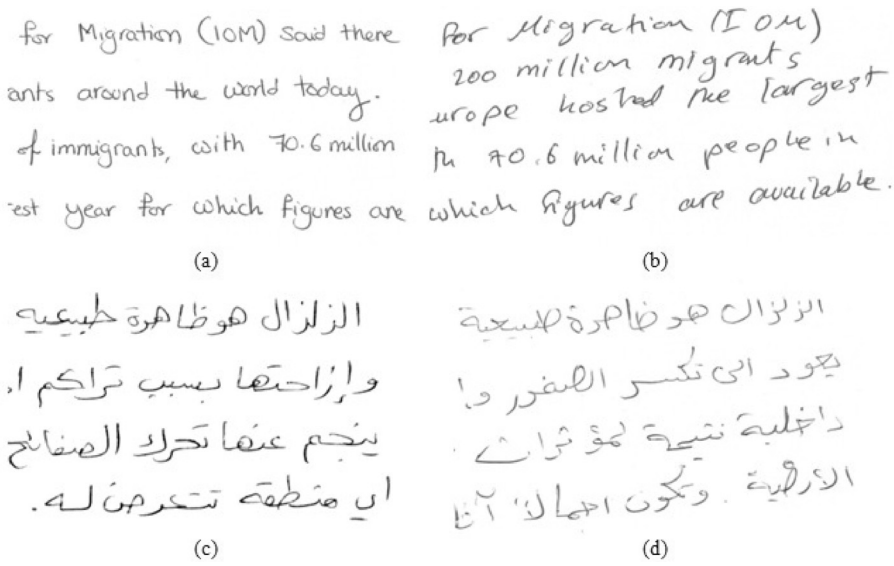

Fig. 5. Writing samples from the QUWI database (a): English female (b): English male (c): Arabic female (d): Arabic male.

Table 3
Performance (in terms of classification rates) of individual classifiers for Scenario-A and Scenario-B.

Learner	Scenario A					Scenario B				
	LBP	HoG	GLCM	SFTA	All	LBP	HoG	GLCM	SFTA	All
ANN	**74**	62	77	**57**	**76**	**74**	**72**	**79**	**55**	**84**
SVM	67	57	**78**	55	68	**74**	66	70	48	78
DT	67	57	67	49	69	58	50	60	48	72
KNN	67	**64**	60	56	58	68	60	63	50	54
RF	63	55	67	50	62	57	50	51	49	61

Table 4
Performance (in terms of classification rates) of individual classifiers for Scenario-C and Scenario-D.

Learner	Scenario C					Scenario D				
	LBP	HoG	GLCM	SFTA	All	LBP	HoG	GLCM	SFTA	All
ANN	**74**	**67**	**70**	52	**79**	**72**	**69**	**80**	52	**74**
SVM	71	60	**70**	52	75	63	59	75	47	71
DT	66	58	56	49	66	68	57	55	50	64
KNN	65	64	64	**58**	57	58	61	62	47	50
RF	59	58	62	50	63	64	50	62	**57**	64

well their combination are presented in Tables 5–8 for the four scenarios. In general, the classification rates increase when using a combination of classifiers for each of the individual features and their combination. For Scenario-A, the classification rates increase from 74%, 64%, 78% and 57% to 79%, 66%, 79% and 60% for LBP, HoG, GLCM and SFTA features, respectively. The classification rate of the feature combination is 76% in both cases. Likewise, for English handwriting samples (Scenario-B) the classification rates rise from 74%, 72%, 79%, 55% and 84% to 75%, 73%, 82%, 60% and 85% for the four features and their combination, respectively.

For script-independent evaluations in Scenarios C and D, a similar trend can be observed. Using Arabic samples in training and English samples in the test set (Scenario-C) the gender classification performance increase from 74%, 67%, 70%, 58% and 79% to 73%, 68%, 73%, 59% and 80% using LBP, HoG, GLCM, SFTA and the combination of all features, respectively. In a similar fashion, an enhanced performance is observed for Scenario-D for all the features and their combination where the classification rate increase from 72%, 69%, 80%, 57% and 74% to 75%, 71%, 81%, 58% and 77%, respectively.

Comparing the performance of different ensemble classifiers, for bagging, BAG-ANN (Bagging with Artificial Neural Network as base learner) and BAG-SVM (Bagging with SVM as base learner) report the best performance of 77% for Scenario A with GLCM features. For the voting classifier, the best combination is of KNN, DT and ANN base learners with LBP features. Stacking meta-learner with SVM and ANN stacked together reports the best performance of 79% with GLCM features. The trend is more or less the same for other scenarios as well where the top performing individual classifiers give the best combination of base learners in the ensemble with up to 85% accuracy. For example, in scenario B, SVM and ANN combined as base learners for stacking and voting meta-learners realize 85% and 80% classification rates (using all features) respectively. In stacking meta learner, DT comes out to be the best stacked learner at level 1 as compared to KNN.

It can be concluded that the learners that perform well individually prove to be the best combination of base learners in meta-learning techniques as well. Like in our case, ANN and SVM come out to be the best individual classifiers as well as good base learners. Comparing different ensemble classifiers, though different

Table 5
Performance (in terms of classification rates) of ensemble classifiers for Scenario-A.

Meta learner	Base learner		LBP	HoG	GLCM	SFTA	All
Bagging	ANN		74	63	75	50	74
	SVM		69	54	77	55	68
	DT		71	62	67	46	57
	KNN		67	62	59	55	55
	RF		61	55	70	60	68
Voting	KNN, SVM, ANN		69	63	77	56	69
	KNN, SVM, DT		70	63	72	55	69
	SVM, DT, ANN		74	58	78	59	75
	KNN, DT, ANN		**79**	**66**	73	**60**	73
	KNN, SVM, ANN, DT		68	63	72	56	69
Stacking	Level 0	Level 1					
	SVM + KNN	DT	65	57	71	49	69
	ANN+SVM	DT	74	63	78	49	**76**
	SVM+ANN	KNN	67	63	**79**	56	58
	SVM+DT	KNN	66	61	**79**	56	58
	SVM+KNN+DT	KNN	66	65	73	58	58
	SVM+ANN+KNN	DT	74	63	71	49	**76**

Table 6
Performance (in terms of classification rates) of ensemble classifiers for Scenario-B.

Meta learner	Base learner		LBP	HoG	GLCM	SFTA	All
Bagging	ANN		73	72	**82**	54	84
	SVM		74	67	72	48	78
	DT		67	61	70	52	72
	KNN		68	61	65	51	54
	RF		66	68	73	**60**	71
Voting	KNN, SVM, ANN		**75**	67	78	52	80
	KNN, SVM, DT		72	68	70	45	77
	SVM, DT, ANN		71	70	81	51	77
	KNN, DT, ANN		72	65	71	48	80
	KNN, SVM, ANN, DT		74	68	73	52	77
Stacking	Level 0	Level 1					
	SVM + KNN	DT	66	65	70	48	76
	ANN+SVM	DT	74	72	70	48	84
	SVM+ANN	KNN	69	72	71	50	54
	SVM+DT	KNN	69	66	68	50	54
	SVM+KNN+DT	KNN	69	67	70	51	54
	SVM+ANN+KNN	DT	74	**73**	70	48	**85**

Table 7
Performance (in terms of classification rates) of ensemble classifiers for Scenario-C.

Meta learner	Base learner		LBP	HoG	GLCM	SFTA	All
Bagging	ANN		69	64	49	47	76
	SVM		71	61	70	55	75
	DT		66	55	59	55	70
	KNN		65	62	62	56	50
	RF		71	65	66	51	75
Voting	KNN, SVM, ANN		68	64	**73**	55	76
	KNN, SVM, DT		69	61	66	56	67
	SVM, DT, ANN		**75**	65	72	52	**80**
	KNN, DT, ANN		68	64	67	53	69
	KNN, SVM, ANN, DT		68	65	72	55	72
Stacking	Level 0	Level 1					
	SVM + KNN	DT	69	58	70	49	77
	ANN+SVM	DT	68	67	70	49	79
	SVM+ANN	KNN	65	67	71	58	57
	SVM+DT	KNN	65	57	69	58	57
	SVM+KNN+DT	KNN	65	60	71	58	57
	SVM+ANN+KNN	DT	70	**68**	70	**59**	79

combinations of classifiers work differently with different features in different scenarios, in general, voting appears to be the most effective of the ensembles realizing highest classification rates in many cases as compared to bagging and stacking (Tables 5–8).

Analyzing the classification rates across the four scenarios, it can be seen that highest classification rates of 79%, 85%, 80% and 81% are realized for scenarios A, B, C and D, respectively. The confusion matrices corresponding to the best performing classifier for each of the scenarios are illustrated in Fig. 6. It is interesting to note that for scenarios A and B, the errors are more or less equally distributed among male and female writers. For scenarios C and D, on the other hand, it can be seen that the number of male writers falsely classified as females is almost double as compared to the number of female writers recognized as males. It seems like

Table 8
Performance (in terms of classification rates) of ensemble classifiers for Scenario-D.

Meta learner	Base learner		LBP	HoG	GLCM	SFTA	All
Bagging	ANN		70	50	74	46	75
	SVM		63	60	78	47	71
	DT		63	65	71	52	75
	KNN		64	56	63	**58**	50
	RF		71	60	76	50	74
Voting	KNN, SVM, ANN		70	65	**81**	47	74
	KNN, SVM, DT		64	61	70	47	62
	SVM, DT, ANN		67	65	79	48	73
	KNN, DT, ANN		71	67	75	51	65
	KNN, SVM, ANN, DT		70	65	74	48	**77**
Stacking	Level 0	Level 1					
	SVM + KNN	DT	68	61	75	47	74
	ANN+SVM	DT	**75**	70	75	50	74
	SVM+ANN	KNN	59	70	77	47	50
	SVM+DT	KNN	59	60	74	47	50
	SVM+KNN+DT	KNN	59	60	75	47	50
	SVM+ANN+KNN	DT	**75**	**71**	75	47	74

	True:Female	True:Male
Predicted:Female	39	10
Perdicted: Male	11	40

Scenario A: 79% (Stacking)

	True:Female	True:Male
Predicted:Female	43	8
Perdicted: Male	7	42

Scenario B: 85% (Stacking)

	True:Female	True:Male
Predicted:Female	44	14
Perdicted: Male	6	36

Scenario C: 80% (Voting)

	True:Female	True:Male
Predicted:Female	43	12
Perdicted: Male	7	38

Scenario D: 81% (Voting)

Fig. 6. Confusion matrices corresponding to the best performing classifier in the four scenarios.

Table 9
Results of ICDAR 2015 gender classification competition (Djeddi et al., 2015).

Method	Classification rate (Rank)			
	Task A	Task B	Task C	Task D
LISIC	60(3)	42(8)	49(5)	55(2)
ACIRS	60(3)	54(3)	53(3)	49(6)
Nuremberg	62(2)	**60(1)**	55(2)	53(3)
MCS-NUST	47(7)	51(5)	48(6)	45(8)
CVC	**65(1)**	57(2)	**63(1)**	**58(1)**
QU	44(8)	52(4)	53(3)	47(7)
UBMA	51(5)	50(6)	44(7)	50(5)
ESI-STIC	48(6)	46(7)	42(8)	53(3)
Proposed method (Indv. classifiers)	**78**	**84**	**79**	**80**
Proposed method (Ensemble)	**79**	**85**	**80**	**81**

that female writers are more consistent across varying scripts as opposed to male writers. In our further study, we intend to carry out experiments on larger datasets to validate this hypothesis.

We also compare the performance of our system with those reported in the ICDAR 2015 gender classification competition (Djeddi et al., 2015). As mentioned earlier, for a meaningful comparison, we have employed the same datasets and the evaluation protocol as those of the competition. The classification rates of systems submitted to the competition are summarized in Table 9. Against each classification rate, the number in the parenthesis represents the relative rank (position) of the submitted system in the competition. We present the results of our system for the individual classifiers as well as for the ensemble. It can be seen from Table 9 that the proposed system realizes better classification rates than those of the winner systems in the competition under exactly the same experimental settings. Not only the ensemble classifier, but the individual classifier results also outperform the state-of-the-art results on this problem.

5. Conclusion

We presented an effective technique to determine gender of the writer of a handwritten sample. The technique mainly relies on extracting a set of textural features and training multiple classifiers using writing samples of male and female writers. Different classifiers are then combined using bagging, voting and stacking. The system was evaluated on writing samples of the QUWI database using the same experimental setup as that of the ICDAR 2015 gender classification competition. A series of evaluations revealed the superiority of the proposed technique over existing systems realizing high classification rates on this challenging problem. The interesting scenario of script independent evaluations was also considered in our experimental study where training and test writing samples come from different scripts.

In our further study on this subject, we intend to work on the front-end design of experiments for tuning the parameters of the various classifiers used. We also plan to extend the system to online handwriting samples where additional information in terms of writing speed, pen pressure and time taken etc. is also available. In some cases in the present system, an individual feature reports higher classification rates as compared to the feature combination. This suggests carrying out a feature selection study to find the optimal combination of features for this problem. We also plan to evaluate the system on the complete QUWI database to study how the system performance evolves as a function of the database size. Likewise, in addition to gender, prediction of other attributes like

age, handedness and race etc. will also make the subject of our further study on this interesting problem.

References

Acton, Q. A. (2013). *Advances in machine learning research and application*. Scholarly Editions.

Al Máadeed, S., Ayouby, W., Hassaine, A., & Aljaam, J. M. (2012). QUWI: An Arabic and English handwriting dataset for offline writer identification. In *Proceedings of the international conference on frontiers in handwriting recognition (ICFHR)* (pp. 746–751). IEEE.

Al-Maadeed, S., Ferjani, F., Elloumi, S., & Jaoua, A. (2016). A novel approach for handedness detection from off-line handwriting using fuzzy conceptual reduction. *EURASIP Journal on Image and Video Processing, 2016*(1), 1–14.

Al Maadeed, S., & Hassaine, A. (2014). Automatic prediction of age, gender, and nationality in offline handwriting. *EURASIP Journal on Image and Video Processing, 1*, 1–10.

Alpaydin, E. (2014). *Introduction to machine learning*. MIT Press.

Bandi, K. R., & Srihari, S. N. (2005). Writer demographic classification using bagging and boosting. In *Proceedings of the twelvth international graphonomics society conference* (pp. 133–137).

Beech, J., & Mackintosh, I. (2005). Do differences in sex hormones affect handwriting style? Evidence from digit ratio and sex role identity as determinants of the sex of handwriting. *Personality and Individual Differences, 39*(2), 459–468.

Bertolini, D., Oliveira, L. S., Justino, E., & Sabourin, R. (2013). Texture-based descriptors for writer identification and verification. *Expert Systems with Applications, 40*(6), 2069–2080.

Bouadjenek, N., Nemmour, H., & Chibani, Y. (2014). Local descriptors to improve off-line handwriting-based gender prediction. In *Proceedings of the sixth international conference on soft computing and pattern recognition (SOCPAR)* (pp. 43–47).

Bouadjenek, N., Nemmour, H., & Chibani, Y. (2015a). Age, gender and handedness prediction from handwriting using gradient features. In *Proceedings of the thirteenth international conference on document analysis and recognition (ICDAR)* (pp. 1116–1120). IEEE.

Bouadjenek, N., Nemmour, H., & Chibani, Y. (2015b). Histogram of oriented gradients for writer's gender, handedness and age prediction. In *Proceedings of the international symposium on innovations in intelligent systems and applications (INISTA)* (pp. 1–5).

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123–140.

Brink, A., Smit, J., Bulacu, M., & Schomaker, L. (2012). Writer identification using directional ink-trace width measurements. *Pattern Recognition, 45*(1), 162–171.

Buitrago, M. M., Ringer, T., Schulz, J. B., Dichgans, J., & Luft, A. R. (2004). Characterization of motor skill and instrumental learning time scales in a skilled reaching task in rat. *Behavioural Brain Research, 155*(2), 249–256.

Caligiuri, M. P., & Mohammed, L. A. (2012). *The neuroscience of handwriting: Applications for forensic document examination*. CRC Press.

Cha, S.-H., & Srihari, S. (2001). A priori algorithm for sub-category classification analysis of handwriting. In *Proceedings of the sixth international conference on document analysis and recognition* (pp. 1022–1025).

Cohen, M. R. (1997). Individual and sex differences in speed of handwriting among high school students. *Perceptual and Motor Skills, 84*, 1428–1430.

Costa, A. F., Humpire-Mamani, G., & Traina, A. J. M. (2012). An efficient algorithm for fractal analysis of textures. In *Proceedings of the twenty-fifth SIBGRAPI conference on graphics, patterns and images* (pp. 39–46). IEEE.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition: 1* (pp. 886–893). IEEE.

Déniz, O., Bueno, G., Salido, J., & De la Torre, F. (2011). Face recognition using histograms of oriented gradients. *Pattern Recognition Letters, 32*(12), 1598–1603.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Multiple classifier systems* (pp. 1–15). Springer.

Djeddi, C., Al-Maadeed, S., Gattal, A., Siddiqi, I., Souici-Meslati, L., & El Abed, H. (2015). ICDAR2015 competition on multi-script writer identification and gender classification using quwidatabase. In *Proceedings of the thirteenth international conference on document analysis and recognition (ICDAR)* (pp. 1191–1195). IEEE.

Dorfberger, S., Adi-Japha, E., & Karni, A. (2009). Sex differences in motor performance and motor learning in children and adolescents: an increasing male advantage in motor learning and consolidation phase gains. *Behavioural Brain Research, 198*(1), 165–171.

Džeroski, S., & Ženko, B. (2004). Is combining classifiers with stacking better than selecting the best one? *Machine Learning, 54*(3), 255–273.

Ebrahimzadeh, R., & Jampour, M. (2014). Efficient handwritten digit recognition based on histogram of oriented gradients and SVM. *International Journal of Computer Applications, 104*(9), 10–13.

Feder, K. P., & Majnemer, A. (2007). Handwriting development, competency, and intervention. *Developmental Medicine & Child Neurology, 49*(4), 312–317.

Fiel, S., & Sablatnig, R. (2013). Writer identification and writer retrieval using the fisher vector on visual vocabularies. In *Proceedings of the twelvth international conference on document analysis and recognition (ICDAR)* (pp. 545–549).

Francks, C., DeLisi, L. E., Fisher, S. E., Laval, S. H., Rue, J. E., Stein, J. F., et al. (2003). Confirmatory evidence for linkage of relative hand skill to 2p12-q11. *American Journal of Human Genetics, 72*(2), 499–502.

Fuentes, C. T., Mostofsky, S. H., & Bastian, A. J. (2009). Children with autism show specific handwriting impairments. *Neurology, 73*(19), 1532–1537.

Goodenough, F. (1945). Sex differences in judging the sex of handwriting. *Journal of Social Psychology, 22*, 61–68.

Gordo, A., Fornés, A., & Valveny, E. (2013). Writer identification in handwritten musical scores with bags of notes. *Pattern Recognition, 46*(5), 1337–1345.

Gulcher, J. R., Jónsson, P., Kong, A., Kristjánsson, K., Frigge, M. L., Kárason, A., … Sigurdardottir, S., et al. (1997). Mapping of a familial essential tremor gene, fet1, to chromosome 3q13. *Nature Genetics, 17*(1), 84–87.

Hamid, S., & Loewenthal, K. (1996). Inferring gender from handwriting in Urdu and English. *Journal of Social Psychology, 136*(6), 778–782.

Hartley, J. (1991). Sex differences in handwriting: A comment on spear. *British Educational Research Journal, 17*(2), 141–145.

Hayes, W. N. (1996). Identifying sex from handwriting. *Perceptual and Motor Skills, 83*, 91-800.

Ho, T. K., Hull, J. J., & Srihari, S. N. (1994). Decision combination in multiple classifier systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 16*(1), 66–75.

Joutel, G., Eglin, V., Bres, S., & Emptoz, H. (2007). Curvelets based queries for CBIR application in handwriting collections. In *Proceedings of the ninth international conference on document analysis and recognition: 2* (pp. 649–653). IEEE.

Kobayashi, T., Hidaka, A., & Kurita, T. (2008). Selection of histograms of oriented gradients features for pedestrian detection. In *Neural information processing* (pp. 598–607). Springer.

Koppenhaver, K. M. (2007). *Factors that cause changes in handwriting, forensic document examination: principles and practice* (pp. 27–36). Humana Press.

Kuncheva, L. I. (2004). *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons.

Kushki, A., Chau, T., & Anagnostou, E. (2011). Handwriting difficulties in children with autism spectrum disorders: A scoping review. *Journal of Autism and Developmental Disorders, 41*(12), 1706–1716.

Liatsis, P. (2002). *Recent trends in multimedia information processing*. World Scientific.

Liwicki, M., Schlapbach, A., & Bunke, H. (2011). Automatic gender detection using on-line and off-line information. *Pattern Analysis and Applications, 14*(1), 87–92.

Marsland, S. (2014). *Machine learning: an algorithmic perspective*. CRC Press.

Minetto, R., Thome, N., Cord, M., Leite, N. J., & Stolfi, J. (2013). T-hog: An effective gradient-based descriptor for single line text regions. *Pattern Recognition, 46*(3), 1078–1090.

Ojala, T., Pietikainen, M., & Harwood, D. (1994). Performance evaluation of texture measures with classification based on kullback discrimination of distributions. In *Proceedings of the twelvth IAPR international conference on pattern recognition, conference a: Computer vision and image processing* (pp. 582–585).

Ojala, T., Pietikainen, M., & Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*(7), 971–987.

Polikar, R., Zhang, C., & Ma, Y. (2012). *Ensemble machine learning: Methods and applications: Ensemble learning* (pp. 1–34). Springer-Verlag.

Rosenblum, S., Engel-Yeger, B., & Fogel, Y. (2013). Age-related changes in executive control and their relationships with activity performance in handwriting. *Human Movement Science, 32*(2), 363–376.

Rybski, P. E., Huber, D., Morris, D. D., & Hoffman, R. (2010). Visual classification of coarse vehicle orientation using histogram of oriented gradients features. In *Proceedings of the IEEE intelligent vehicles symposium (iv)* (pp. 921–928). IEEE.

Sammut, C., & Webb, G. I. (2011). *Encyclopedia of machine learning*. Springer Science & Business Media.

Schomaker, L. (2008). *Writer identification and verification* (pp. 247–264). London: Springer.

Schröter, A., Mergl, R., Bürger, K., Hampel, H., Möller, H.-J., & Hegerl, U. (2003). Kinematic analysis of handwriting movements in patients with Alzheimers disease, mild cognitive impairment, depression and healthy subjects. *Dementia and Geriatric Cognitive Disorders, 15*(3), 132–142.

Sharma, N., Pal, U., & Blumenstein, M. (2014). A study on word-level multi-script identification from video frames. In *Proceedings of the international joint conference on neural networks (IJCNN)* (pp. 1827–1833). IEEE.

Siddiqi, I., Cloppet, F., & Vincent, N. (2009). Contour based features for the classification of ancient manuscripts. In *Proceedings of the conference of the international graphonomics society* (pp. 226–229).

Siddiqi, I., Djeddi, C., Raza, A., & Souici-meslati, L. (2014). Automatic analysis of handwriting for gender classification. *Pattern Analysis and Applications*, 1–13.

Sokic, E., Salihbegovic, A., & Ahic-Djokic, M. (2012). Analysis of off-line handwritten text samples of different gender using shape descriptors. In *Proceedings of the ninth international symposium on telecommunications (BIHTEL)* (pp. 1–6).

Teulings, H.-L., & Stelmach, G. E. (1991). Control of stroke size, peak acceleration, and stroke duration in parkinsonian handwriting. *Human Movement Science, 10*(2), 315–334.

Ting, K. M., & Witten, I. H. (1997). Stacked generalization: when does it work? In *Proceedings of the international joint conference on artificial intelligence* (pp. 866–871).

Traina Jr, C., Traina, A., Wu, L., & Faloutsos, C. (2010). Fast feature selection using fractal dimension. *Journal of Information and data Management, 1*(1), 3.

Van Galen, G. P. (1991). Handwriting: Issues for a psychomotor theory. *Human Movement Science, 10*(2), 165–191.

Van Galen, G. P., Van Doorn, R. R., & Schomaker, L. R. (1990). Effects of motor programming on the power spectral density function of finger and wrist movements. *Journal of Experimental Psychology: Human Perception and Performance, 16*(4), 755.

Walton, J. (1997). Handwriting changes due to aging and Parkinson's syndrome. *Forensic Science International, 88*(3), 197–214.

Wang, X., Han, T. X., & Yan, S. (2009). An HOG-LBP human detector with partial occlusion handling. In *Proceedings of the IEEE twelvth international conference on computer vision* (pp. 32–39). IEEE.

Weintraub, N., Drory-Asayag, A., Dekel, R., Jokobovits, H., & Parush, S. (2007). Developmental trends in handwriting performance among middle school children. *OTJR: Occupation, Participation and Health, 27*(3), 104–112.

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks, 5*(2), 241–259.

Woods, K., Bowyer, K., & Kegelmeyer Jr, W. P. (1996). Combination of multiple classifiers using local accuracy estimates. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition* (pp. 391–396). IEEE.

Wu, X., Tang, Y., & Bu, W. (2014). Offline text-independent writer identification based on scale invariant feature transform. *IEEE Transactions on Information Forensics and Security, 9*(3), 526–536.

Xiong, Y.-J., Wen, Y., Wang, P. S., & Lu, Y. (2015). Text-independent writer identification using sift descriptor and contour-directional feature. In *Proceedings of the thirteenth international conference on document analysis and recognition (ICDAR)* (pp. 91–95).

Yosef, I. B., Kedem, K., Dinstein, I. h., Beit-Arie, M., & Engel, E. (2004). Classification of Hebrew calligraphic handwriting styles: preliminary results. In *Proceedings of the first international workshop on document image analysis for libraries* (pp. 299–305). IEEE.

Zhang, D., & Tsai, J. J. (2005). *Machine learning applications in software engineering*: 16. World Scientific.