Pattern Anal Applic (2015) 18:887–899
DOI 10.1007/s10044-014-0371-0

SHORT PAPER

Automatic analysis of handwriting for gender classification

Imran Siddiqi · Chawki Djeddi · Ahsen Raza ·
Labiba Souici-meslati

Received: 19 October 2013 / Accepted: 15 April 2014 / Published online: 6 May 2014
© Springer-Verlag London 2014

Abstract This paper presents a study to predict gender of individuals from scanned images of their handwritings. The proposed methodology is based on extracting a set of features from writing samples of male and female writers and training classifiers to learn to discriminate between the two. Writing attributes like slant, curvature, texture and legibility are estimated by computing local and global features. Classification is carried out using artificial neural networks and support vector machine. The proposed technique evaluated on two databases under a number of scenarios realized interesting results on predicting gender from handwriting.

Keywords Gender prediction · Handwriting · Neural networks · SVM

I. Siddiqi (✉)
Department of Computer Science, Bahria University, Islamabad, Pakistan
e-mail: imran.siddiqi@bahria.edu.pk

C. Djeddi
LAMIS Laboratory, University of Tebessa, Tebessa, Algeria
e-mail: c.djeddi@mail.univ-tebessa.dz

C. Djeddi
Department of Computer Science, Badji Mokhtar-Annaba University, Annaba, Algeria

A. Raza
Medical Transcription Billing Corporation, Islamabad, Pakistan
e-mail: ehsanraza@mtbc.com

L. Souici-meslati
LISCO Laboratory, Badji Mokhtar-Annaba University, Annaba, Algeria
e-mail: labiba.souici@univ-annaba.org

1 Introduction

Handwriting is one of the oldest modes of communication in our civilization which has developed and evolved over time. An individual learns to write by copying shapes from a standard copy book which itself varies as a function of the geographical location, temporal, social and cultural circumstances. The learned copy book style, however, diminishes with time and an individual develops his/her own writing preferences. In this context, as opposed to electronic or printed text, handwritten text carries additional information about the individual who produced the text. This makes analysis of handwriting an attractive research area for psychologists, document examiners, palaeographers, graphologists and forensic analysts. Although a significant number of organizations employ handwriting analysis for personality profiling [25, 42], the correlation between personality and handwriting still remains debatable [15, 26, 34, 49] and is yet to be validated on scientific grounds [3]. The only meaningful correlation that has been experimentally validated exists between handwriting and the gender of the writer [3, 9, 17, 18, 20, 21]. This classification of gender from handwriting has been an interesting research topic since the initial decades of last the century [8, 35, 48] and has matured significantly since then. With the advancements in image analysis and pattern classification techniques, manual analysis of handwriting is being replaced with automated systems.

A large number of systems have been proposed and developed for automatic analysis of handwritten documents mainly targeting applications like handwriting recognition, word spotting, writer identification and signature verification. Automatic classification of gender from handwriting, however, has been a relatively less explored area with only few significant contributions.

Identifying the demographic classes including gender, handedness and age from handwritten documents has been investigated in [2, 11]. A set of global (macro) features like slant, word gap, gray-scale threshold, etc. [47] is computed to discriminate between different demographic classes of writers. Classification is carried out using neural networks which are combined through bagging and boosting reporting classification rates in the range of 75–87 % for different demographic classes.

Hamid and Loewenthal [18] conducted a study using writing samples in English and Urdu and presented 'delicacy and decorativeness' as a major discriminating factor between writings of males and females. A consistent classification rate of about 68 % was achieved on both English and Urdu writing samples.

In [28], authors employ a combination of online and offline features to predict gender from handwriting and report an accuracy of 67.5 % using Gaussian Mixture Models as classifier. The study also claims that online information produces better classification as opposed to the features extracted from offline representation of the online data. The details of the features used in this study can be found in [29].

In a relatively recent study [46], the authors employ Fourier descriptors, tangent and curvature information and bending energy to characterize gender from handwritten samples. The results however are not reported in a quantified form and only the values of different features computed from the same word for male and female writers are presented and discussed.

In this paper, we present a system for automatic classification of gender from handwriting using a set of features aimed at computing a subset of the discriminative attributes identified by the psychologists. Each writing sample is represented by a set of features which is fed to a classifier to learn to distinguish between the two classes: male and female. Classification is carried out using neural networks and support vector machines. The proposed method evaluated on two datasets reports interesting results.

The main contribution of this paper is the analysis of different types of features capturing the orientation, curvature and legibility information in the writing for predicting the gender of the writer. In addition to an analysis of the performance of these different types of features, experiments on two totally different datasets comprising writing samples of text in different languages (Arabic, English and French) allow interesting analyses of classification performance in scenarios including text-independent and text-dependent, and script-dependent and script-independent modes. The impact of geographical location of writers on gender classification is also studied through cross-database evaluations. The effectiveness of the proposed system is evidenced by the promising classification rates obtained on a much larger database of writing samples as compared to those employed by the state-of-the-art methods on this subject.

This paper is organized as follows. In the next section, we discuss some attributes of handwriting which serve to distinguish writings of male and female writers. We then present the proposed set of features in Sect. 3 followed by the classification scheme in Sect. 4. Experimental results and their analysis are presented in Sect. 5 while the last section concludes the paper with some discussion on potential future research directions on the subject.

2 Gender differences in handwriting

As discussed earlier, several studies have shown that gender can be predicted from handwriting [18, 20, 21] with varying degrees of success. This is supported by the observation that individuals interacting with handwritten documents, for example, teachers and clerks, learn to discriminate between male and female writings with time. Untrained human examiners are also able to predict gender from writing above-chance level [46]. Psychologists attribute the differences in handwriting of males and females to differences in motor coordination [20] or the different types of hormones they produce [21]. Whatever the case be, the researchers do agree on correlation between gender and handwriting.

Typically, the psychologists suggest that attributes like neat, even, well-organized, rounded, small and symmetrical are characteristics of a female writing. On the other hand, hurried, uneven, messy, spiky and sloping writings are most likely to belong to a male writer [9, 20]. Some examples of male and female writing samples supporting these findings have been illustrated in Fig. 1. Document examiners have identified a set of 21 discriminating features (qualitative as well as quantitative) which can be effectively employed for analysis of handwriting [23]. These features are generally termed as conventional features [47]. A subset of these features which can be computed algorithmically from scanned images of writing are known as computational features. Examples of computational features include slant, inter- and intra-word spacing, baseline alignment, pen pressure, gradient information, etc. Some of these features have also been successfully applied to gender discrimination [2].

The task of gender classification is closely related to that of writer identification, the difference being that writer identification is a N class while gender classification is a two-class problem. The features that have been effectively applied to writer identification are therefore likely to perform well for gender classification as well [28]. We will

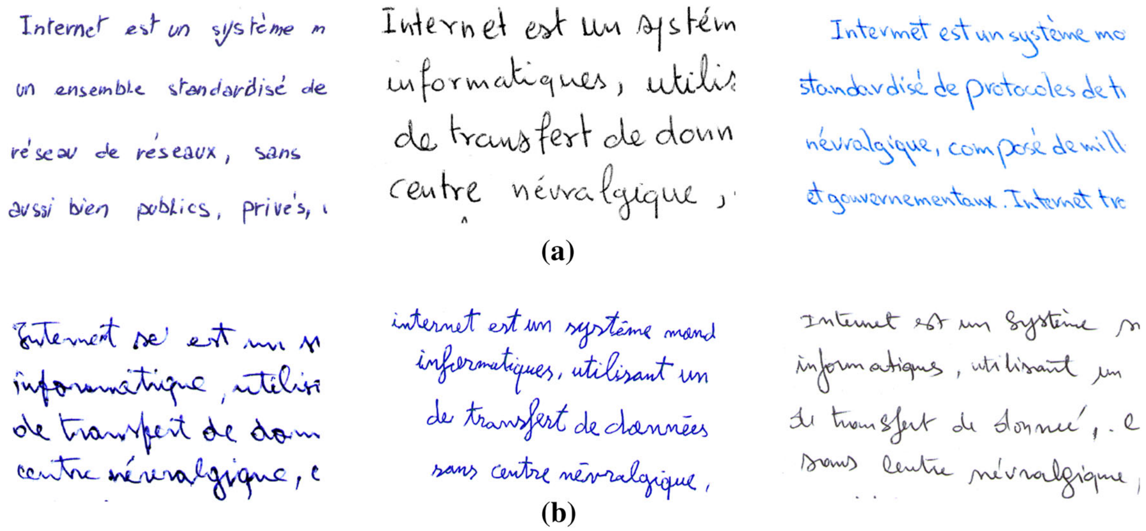


Fig. 1 Sample writings of **a** female writers and **b** male writers

Table 1 Distribution of QUWI database

Data set	Training	Validation	Test
Writers	300	75	100
Samples	1,200	300	400

discuss more about features in Sect. 4, but prior to that we present the databases used in our study in the next section.

3 Datasets

In our study, we have used writing samples from two datasets, the Qatar University Writer Identification (QUWI) [1] database and a custom-developed Multiscript Handwritten Database (MSHD).

In the QUWI database, each writer contributed four pages to the dataset: two in Arabic and two in English. The first and third page of each writer contains an arbitrary text of writer's own choice in Arabic and English respectively while the second and fourth page contains the same text copied by all writers. This allows using the dataset in text-dependent as well as text-independent modes. Performance of systems on different scripts (Arabic and English) can also be studied. We have used writing samples of 475 writers (1,900 samples) in our study.

In our experiments, we divide the QUWI database into three parts. Samples of 300 writers are used as training set, 75 writers as validation set while those of 100 writers are used as the test set. This distribution corresponds to approximately 65 % data as training set, 15 % as validation set and 20 % as test set and, is consistent with the recommended distribution in classification and data mining problems [27]. The distribution of writers into training, validation and test datasets is summarized in Table 1. The distribution of writers stays the same in different experiments. The distribution of samples naturally varies from experiment to experiment as will be discussed in Sect. 6.

The MSHD database comprises writing samples of 87 different writers in French and Arabic. Each writer contributed 12 pages, 6 in French and 6 in Arabic making a total of 1,044 writing samples. The text on sample i of each writer is the same ($i = 1, 2, \ldots, 12$). Gender information of three writers was not available so writings of 84 writers are considered in our evaluations. For most of the evaluations, writings of 42 writers comprise the training set and those of other 42 writers make the test set. The same database has previously been used to evaluate writer identification performance in [14].

The next section discusses the features that we compute from the given writing samples.

4 Feature extraction

Among the different discriminating attributes of male and female writings, we focus on the slant/orientation, roundedness/curvature, neatness/legibility and writing texture in our study. To algorithmically compute these attributes from digitized images of writing, we use a set of features computed at different scales of observation. The slant and curvature are estimated from the contours of writing by representing the contours using: (1) Freeman chain codes and (2) a set of approximating line segments (polygonized contours). Neatness or legibility of writing, although a very subjective attribute, is estimated by computing the fractal dimension of the writing. In previous works [43, 44], orientation- and curvature-based features have shown promising results for tasks of writer identification and verification while fractal dimension has been successfully applied to classification of writings in [7]. Texture-based features are also known to characterize the writer of a document [41]. In this study, we are interested to investigate the effectiveness of these features for the task of gender classification. The extraction of these features is discussed in the following sections.

4.1 Orientation and curvature

The orientation (slant) and curvature information in the writing is extracted by computing a set of features from contours of writing. The contour representation is chosen based on the hypothesis that the shape of characters in a writing can be encapsulated by its contours. Working on contours also eliminates the writing instrument sensitivity while conserving the shape of the characters.

We extract both the interior and exterior contours in writing and as discussed earlier, these contours are represented by a sequence of Freeman chain codes and by a set of polygons obtained by applying a polygonization algorithm to the contours. These two representations correspond to two different scales of observations and the features computed from these different representations complement each other. We discuss extraction of orientation and curvature features from each of these representations in the following sections.

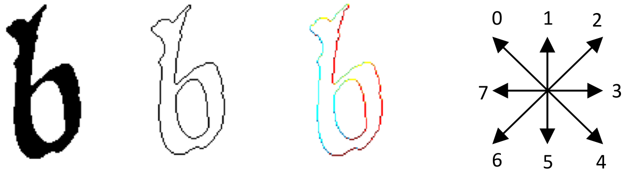


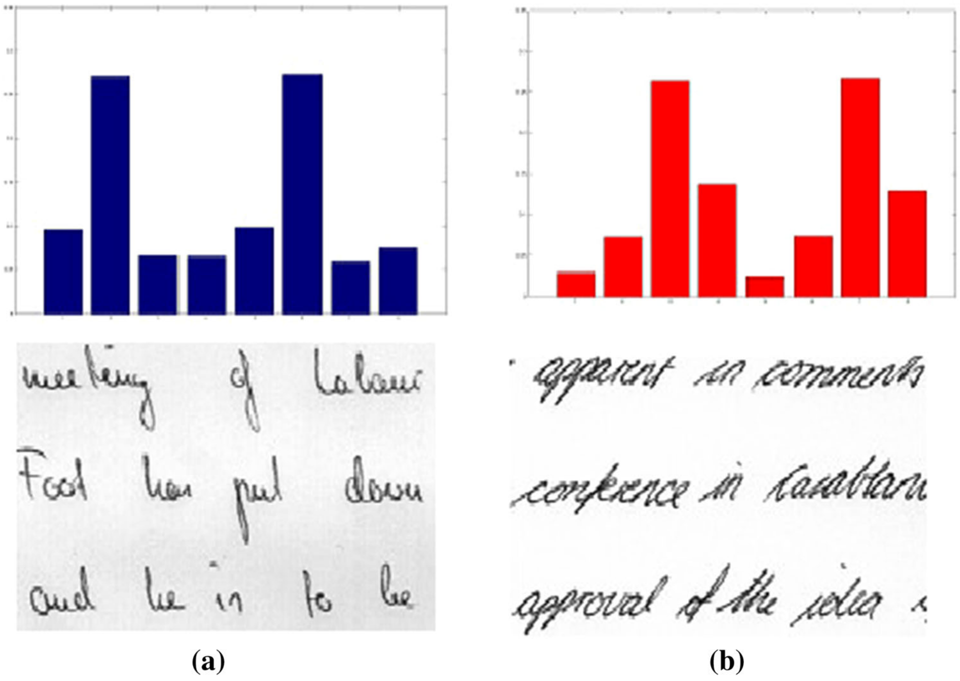
Fig. 2 An image of a character with its contours and chain code representation

4.1.1 Chain code-based features

Chain codes have been effectively applied to problems like character/word recognition [5, 6, 24, 33], classification of writing styles [31] and writer identification [43, 44]. Since our task of gender classification also comprises handwritten documents, we expect chain code representation to be effective for feature extraction. We represent the writing contours by Freeman chain codes. Each contour is a sequence of boundary pixels with $\{c_j | 1 < j \leq M_{i-1}\}$ where $c_j \in \{0, 1, \ldots, 7\}$ and M_i is the length of contour i. An example character with its contour and the codes associated with each of the directions are illustrated in Fig. 2.

Once an image of writing is represented by chain codes, we compute the (normalized) histogram of chain codes, generally termed as slope density function $(f1)$. The (eight) bins of the histogram represent the relative contribution of each of the eight principal directions in a writing while the dominant orientations in the writing are represented as peaks in the histogram. However, it is important to note that since the images are offline, we cannot discriminate

Fig. 3 Writing samples and their respective slope density functions



between forward and backward strokes and the sequence of codes assigned to a particular contour is dependent upon the way it is traversed. A solution could be to quantize the histogram into four bins representing the four principal stroke directions: horizontal, vertical, left diagonal and right diagonal. Our experience, however, has shown that keeping the eight bins and being always consistent in the way a contour is traced is a better choice. Figure 3 illustrates the distribution of chain codes computed from two writing samples. It can be seen that the overall vertical orientation in sample ‘a’ is reflected by two peaks at the respective bins of the corresponding histogram. Similarly, for sample ‘b’ where the writing is tilted towards the right, peaks can be observed at the bins corresponding to the right-diagonal directions.

To estimate curvature at pixel level, we compute the histogram of chain code pairs. We initialize a (8×8) matrix with all bins set to zero. For each pair (i,j) in the chain code representation of a writing, we increment the respective bin of the matrix (histogram). The distribution is finally normalized to be independent of the amount of text. This distribution $(f2)$ could be viewed as a measure of the angle (curvature) between the vectors representing the chain code directions as illustrated in Fig. 4.

The chain code-based computation of orientation and curvature is effective, but since these values are calculated at pixel level they might be sensitive to noise distortions in the writing. To complement these features, we compute similar features by first estimating the contours by a set of polygons. This, in fact, corresponds to a distant scale of observation and the computed features are also robust to noise. These features are discussed in the following subsection.

4.1.2 Polygon-based features

Using the sequential polygonization algorithm in [52], we estimate the contours in writing by a set of line segments (polygons). An example of polygonized contours is illustrated in Fig. 5.

For each segment in the writing we compute its slope and use the distribution of these slopes as our next feature $f3$. The interval $-90°$ to $90°$ is quantized into 8 bins and the slopes of lines approximating the writing contours are counted in their respective bins. The histogram is finally normalized and is used as a feature. To estimate curvature, we compute the angle between each pair of connected segments as

$$\alpha_i = \pi - \arccos \frac{V_i . V_{i+1}}{|V_i| V_{i+1}|} \tag{1}$$

with V_i and V_{i+1} being the vectors from (x_{i-1}, y_{i-1}) to (x_i, y_i) and from (x_i, y_i) to (x_{i+1}, y_{i+1}) respectively as illustrated in Fig. 6.

Similar to the distribution of slopes, the angles $(0° - 180°)$ are quantized into 8 bins and their distribution $f4$ is used to characterize the writing. The implementation details of these features can be found in [44].

After having presented the orientation and curvature features computed from two different scales of observations, we discuss the features based on fractal dimension in the next section.

4.2 Fractal features

The fractal behavior of handwriting was first proved by Vincent [51]. Later studies revealed the effectiveness of fractal features for writer characterization [40] as well as classification of writings [7]. Authors in [7] compute a legibility graph from fractal features and group writings into clusters as a function of their legibility. Fractal features have also been applied to writer identification with acceptable success rates on small data sets [12]. With the aim to capture the regularity and legibility of writing, we chose to compute fractal dimension of the writings under

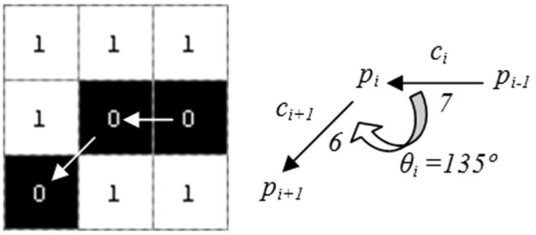
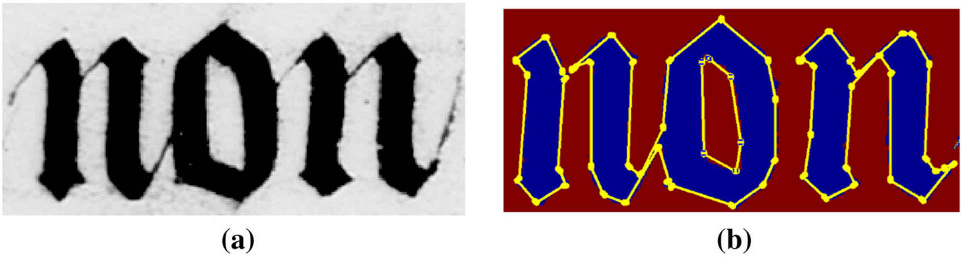


Fig. 4 The chain code pair $(7, 6)$ representing an angle of 135° at pixel position p_i


(a)

(b)

Fig. 5 Polyogonization: **a** original image, **b** polygonized contours

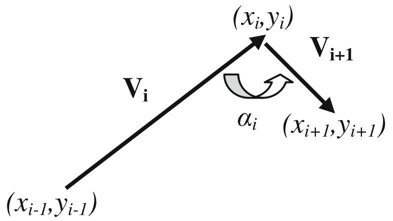


Fig. 6 Angle between neighboring segments of polygonized contours

study and eventually employ it as a feature for discriminating writing samples of male and female writers.

Fractal dimension can be estimated by several methods and a detailed survey of these methods can be found in [30]. Popular categories of these methods include box counting methods, Fractional Brownian motion methods and area measurement methods. In our implementation, we employ the most well-known box counting method to compute the fractal dimension of a given handwritten text.

The basic idea in box counting method is to divide the object (writing in our case) into a number of boxes of size r and counting the number of boxes containing information. The process is repeated by varying the box size and the fractal dimension is estimated as defined in Eq. 2.

$$D = \lim_{r \to 0} \frac{\log N(r)}{\log \frac{1}{r}} \tag{2}$$

where $N(r)$ is the number of boxes of size r needed to cover the object.

The fractal dimension is a single real number and a single value may not be discriminative enough to be used as a feature for a complex problem like gender classification. We, therefore, also introduce multi-fractal analysis and compute the generalized fractal dimensions D_q as a function of moment orders q [50]. The computation of D_q relies on randomly chosing N points belonging to the object and counting for each point i the number of pixels $M_i(r)$ inside boxes of size r. The generalized dimensions D_q are computed using the mean of $M(r)$ for different values of r [30]. In our implementation, we compute the generalized dimensions D_q for $q = 1, 2, 3, 5, 10$ and combine them with the box counting-based fractal dimension to have a six-dimensional feature vector estimating the regularity and legibility of writing.

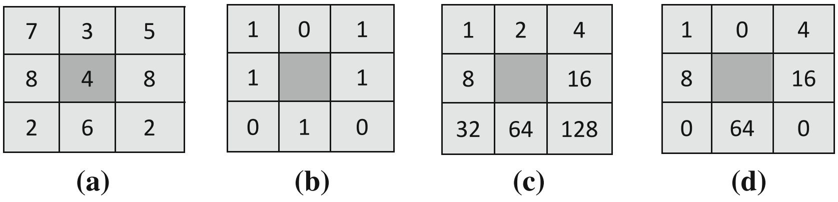
In the next section, we discuss the third category of features, the texture-based features.

4.3 Texture-based features

Texture analysis of handwriting considers each writing as a visually distinctive texture. Texture is related to the overall look and feel of the writing and can be represented in a number of ways. Among the significant texture-based analyses of handwriting, Said et al. [41] employed multi-channel Gabor filters and Gray Level Co-occurrence Matrices (GLCM) to propose a texture-based solution to the writer identification problem. Some recent studies also used these texture-based features for writer identification [45] and verification [19]. Other measures of texture applied to handwriting include local binary patterns (LBP) [4] and auto regressive (AR) coefficients [16]. The performance of these descriptors on tasks like writer identification was found to be better than that of conventional GLCM or Gabor features. We, therefore, chose to employ LBP and AR Coefficients as texture descriptors for possible discrimination between male and female writings. These features are discussed in the following.

4.3.1 Local binary patterns

Local binary patterns were first introduced by Ojala [37, 38] and have been very effectively applied to a number of texture classification applications since then [4, 22, 53]. The original LBP method proposed in [37, 38] consists in generating a limited number of texture units. Considering a set of neighborhood pixels $V = \{V_0, V_1, \ldots, V_8\}$, the adjacent pixels are compared to the central pixel V_0 to generate a binary pattern. The binary assignment is performed as follows. For $i = 1, 2, \ldots, 8$ if $V_i < V_0$ we assign the value 0 to the neighboring pixel i, otherwise, it is assigned the value 1. The resulting pattern is considered as a binary number, and multiplying each bit by the respective weight and summing the values together the LBP code for the central pixel is computed. The process is illustrated in Fig. 7. The histogram of LBP codes provides a descriptor to characterize the texture. In 2002, the authors proposed extensions to their original method to include



7	3	5
8	4	8
2	6	2

(a)

1	0	1
1		1
0	1	0

(b)

1	2	4
8		16
32	64	128

(c)

1	0	4
8		16
0	64	0

(d)

LBP Code = 1+4+8+16+64 = 93

Fig. 7 LBP computation: **a** image values, **b** binary codes assignment, **c** weights of neighboring pixels, **d** conversion to decimal

Table 2 Summary of features

Feature category	Feature	Description	Dimension
Slant and Curv.	$f1$	Distribution of chain codes	8
	$f2$	Distribution of chain code pairs	64
	$f3$	Distribution of segment slopes	8
	$f4$	Distribution of curvatures	8
Fractal features	$f5$	Box counting FD	1
	$f6$	Generalized FD	5
Texture features	$f7$	LBP	243
	$f8$	AR coefficients	24
Total:			361

neighborhood of different sizes and capture dominant features at different scales [39].

The authors also introduced the concept of uniform and non-uniform binary patterns based on the transitions between 0s and 1s in the LBP image. A binary LBP code is considered uniform if the number of transitions is less than or equal to 2, the code can then be seen as a circular string. For example, the code 00100100 is not uniform as it contains 4 transitions but the codes 00000000 and 00100000 are uniform as they contain 0 and 2 transitions, respectively. It was also observed that uniform binary patterns account for most of the patterns in the texture images [39].

In our implementation, we compute the LBP from binary images of handwriting. For p neighboring points, we can have a maximum of $p \times (p-1)+2$ uniform patterns. We use a neighborhood of $p=16$ pixels with a total of 242 $(16 \times (16-2)+2)$ possible uniform patterns. The descriptor (histogram of LBP) therefore comprises 242 bins for the uniform patterns and 1 bin for all non-uniform patterns giving a 243-dimensional feature vector.

4.3.2 *AR coefficients*

Two-dimensional (2D) autoregressive (AR) models were first introduced by Deguchi [13] for image representation and texture characterization. Since then they have been successfully applied to texture segmentation [36] and texture modeling [32]. Recently, AR models have been adapted to characterize and identify the authors of handwritten texts [16].

For our task of gender prediction, we characterize a given writing by a set of two-dimensional (2D) autoregressive coefficients extracted from binary images of writing. To estimate these coefficients, each pixel $x_{i,j}$ in the image is predicted by a linear combination of its neighboring pixels.

Table 3 Number of neurons in each layer for different features

Features	Input neurons	Hidden neurons	Output neurons
Slant and Curv.	88	50	2
Texture	267	150	2
Fractal	6	4	2
All features	361	200	2

$$x_{i,j} = \sum_{p,q \in D} \theta_{pq} x_{i-p} y_{j-q} \tag{3}$$

where D represents the neighborhood context which generally is a rectangular window $D = \{(p,q) | -m \leq p \leq m, -n \leq q \leq n, (p,q) \neq (0,0)\}$. θ_{pq} are the AR coefficients while $p \times q$ is the order of the model. The coefficients are estimated by minimizing the squared error between the predicted and the actual values of the pixel. The details of coefficient estimation can be found in [16]. In our implementation, we used a neighborhood of 5×5 pixels thus giving a total of 24 AR coefficients.

Summarizing, a given handwritten sample is represented by three types of features: slant and curvature, fractal- and texture-based features. Table 2 summarizes these features with the dimensionalities of each.

5 Classification

Classification is carried out using two state-of-the-art classifiers, the artificial neural networks (ANN) and the support vector machine (SVM). The classifiers are trained using the three sets of features extracted from the training data set while the different tunable parameters of the two classifiers are empirically determined on the validation data set.

The ANN is a three-layer network: the input layer having the same number of neurons as the dimension of a particular feature set, the output layer has two neurons corresponding to the two classes (male and female) while the number of neurons in the hidden layer is determined as a function of the dimensionality of the input feature vector (using the validation data set). Table 3 summarizes the number of neurons in each of the three layers for slant and curvature, texture and fractal features as well as their combination. Each neuron has a sigmoid transfer function and the networks are trained using the back propagation algorithm with maximum epochs set to 1000.

In addition to ANN, we also use support vector machine (SVM) to classify the gender of the writing in question. SVMs are typically known to address the problems with many other learning algorithms including local minima, over fitting and an inconveniently large number of tunable parameters. For our system, we have employed the one-against-all SVM implemented using the 'SVM and Kernel Methods Matlab toolbox' described in [10]. The SVM is trained on a polynomial kernel function where the bound on the Lagrangian multipliers 'C' is varied from 10 to 10^7 and the conditioning parameter for QP method lambda is varied from 10^{-1} to 10^{-6} to find the best set of parameters for each of the features (on the validation data set).

Features extracted from a writing sample in question are fed to the trained classifier (ANN or SVM). The classifier outputs the predicted gender of the writer of the document. The performance of both these classifiers on different experimental evaluations is discussed in the next section.

6 Experimental results

This section presents the series of experiments that we carried out to evaluate the effectiveness of the proposed features in predicting the gender of the writing in question. The evaluations are conducted on the QUWI and MSHD databases as presented in Sect. 3. We first present the gender classification rates on the complete data sets using the two classifiers (SVM and ANN) and later describe some interesting scenarios.

In all experiments we make sure that there are no writers with samples in both training and test sets. This may cause a document in question to match with another sample of the same writer in the training set, and this match will in fact correspond to writer identification and not gender recognition. Hence in experiments where more than one sample per writer is considered, all samples of a given writer belong to only one of the training or test sets.

Table 4 presents the classification rates on the two data sets. For QUWI, samples of 300 writers are used for training while those of 100 writers (400 samples) for testing. For MSHD database, samples of 42 writers each are used as training and test sets. Classification rates of as good as 68.75 and 73.02 % are achieved on the QUWI and MSHD databases, respectively. These results are comparable with those obtained by the state-of-the-art methods discussed in Sect. 1. It is however interesting to note that we evaluate the proposed system on much larger databases as opposed to existing methods.

Table 4 Classification rates on QUWI and MSHD databases

Features	Data Set			
	QUWI		MSHD	
	SVM (%)	ANN (%)	SVM (%)	ANN (%)
Slant and Curv.	68.75	67.00	72.82	69.25
Texture	59.75	61.50	68.65	64.88
Fractal	61.50	62.50	62.30	61.90
All features	68.75	67.50	73.02	69.44

Comparing the performance of the two classifiers (ANN and SVM), it can be seen from Table 4 that there is not a very significant difference between the two. Among the three types of features, the slant (orientation) and curvature features outperform texture and fractal features on both the databases. Another interesting observation is that combining the three types of features results only in marginal improvements in the overall classification rates. For subsequent experiments, we therefore discuss the results of individual categories of features only.

In addition to the evaluations on the complete data sets, we also analyze the performance of proposed features in a number of specific scenarios including text-dependent and text-independent, script-dependent and script-independent and cross-database evaluations. These are discussed in the following sections.

6.1 Text-dependent vs. text-independent evaluations

These experiments are aimed at studying how the performance of the features vary if the writing samples in the training and test sets contain the same (different) textual content. For QUWI database, we use the page 2 and page 4 of each writer of text-dependent evaluations on Arabic and English texts, respectively. The reason of using these pages is that page 2 of all writers contains the same text in Arabic while the page 4 of each writer comprises the same text in English allowing text-dependent evaluations. As discussed in Sect. 3, Page 2 (Page 4) of 300 writers is used for training while the same page of 100 writers for testing. For text-independent evaluations, we require the textual content of images in training and test sets to be different. Since page 1 of all writers contain an arbitrary text in Arabic while page 3 in English, for text-independent experiments, we consider pages 1 and 3 for Arabic and English text-independent experiments, respectively. The distribution of writers in the training and test sets is the same as in case of text-dependent experiments (300 writers for training and 100 for test.)

In the MSHD database, the first 6 samples of each writer contain the same text in French while the last 6 samples comprise the same text in Arabic for each of the 84 writers. For text-dependent experiments, we evaluate the system 12 times using sample i of first 42 writers in training and the same sample of the last 42 writers in testing with $i = 1, 2, \ldots, 12$. This allows comparison of the same textual content on Arabic and French texts. The average classification rate for $i = 1, 2, \ldots, 6$ represents the performance on text-dependent French samples while the

Table 5 Classification rates of text-dependent and text-independent evaluations on the QUWI and MSHD databases

Data set	Features	Mode			
		Text-dependent		Text-Independent	
		SVM (%)	ANN (%)	SVM (%)	ANN (%)
QUWI-English	Slant and Curv.	68.00	70.00	70.00	66.00
	Texture	63.00	62.00	62.00	61.00
	Fractal	65.00	65.00	65.00	64.00
QUWI-Arabic	Slant and Curv.	69.00	71.00	63.00	62.00
	Texture	65.00	63.00	63.00	63.00
	Fractal	66.00	66.00	62.00	65.00
MSHD-French	Slant and Curv.	68.25	67.06	67.46	66.27
	Texture	66.67	66.27	66.27	65.48
	Fractal	64.68	66.27	63.09	65.87
MSHD-Arabic	Slant and Curv.	73.41	72.62	68.65	69.44
	Texture	74.20	72.22	72.22	71.43
	Fractal	65.08	65.87	64.28	65.08

same value for $i = 7, 8, \ldots, 12$ represents the classification rate on text-dependent Arabic samples. In text-independent evaluations, features extracted from different textual content need to be compared. Therefore, the first 3 French (Arabic) samples of 42 writers are used for training and the last 3 samples of other 42 writers for testing. Later, the last 3 French (Arabic) images are used as training and the first 3 as test set. The overall classification rates for these experiments are computed as the average of the two runs.

It should also be noted that in all the subsets of data discussed above, no writers are common in the training and test sets. The results of these evaluations are summarized in Table 5. In general, the classification rates of text-dependent and text-independent experiments follow the same trends as in Table 4 with slant and curvature features performing better than the texture and fractal features in most of the cases. Comparing the text-independent and text-dependent classification rates, except for one experiment (QUWI-English, SVM classifier), the classification rates of text-dependent evaluations are better (although marginally in most cases) than those of text-independent evaluations. Another very interesting observation is that the slant and curvature features, in general, are more sensitive to the textual content of images as opposed to texture or fractal features. This seems very much natural as slant and curvature represent local features of writing and hence are more sensitive to the image content. The texture and fractal features being global attributes of writing are relatively less sensitive to the content and hence exhibit less variation in the text-dependent and text-independent classification rates.

In the next section, we present the results of script-dependent and script-independent evaluations.

6.2 Script-dependent vs. script-independent evaluations

These experiments are aimed at studying how the classification rates vary if the same/different scripts are used as training and test sets. In script-dependent experiments, the writing samples in the same script (English, French or Arabic) are used both in training and test sets while in script-independent experiments the training and test data sets comprise samples in different scripts.

On the QUWI database, the script-dependent evaluations on Arabic text involve pages 1 and 2 of 300 writers in training and the same pages of 100 writers in testing. For experiments on English text, pages 3 and 4 are used with the same distribution in training and test sets. This makes a total of 600 training and 200 test samples for each (Arabic and English) set of experiments. On the MSHD database, the first 6 samples of each writer which are written in French are used for script-dependent evaluations on French text with 42 writers in training and 42 writers in the test set. Similarly, the last 6 samples of each writer (which are written in Arabic) are used for script-dependent evaluations on Arabic text.

Script-independent experiments are more challenging and involve training samples in a different script than the test samples. For the QUWI database, in the first experiment, the English samples of 300 writers are used in training and Arabic samples of 100 writers in testing. Later, the scenario is reversed by employing Arabic samples of 300 writers for training and the English samples of 100 writers for testing. For experiments with MSHD database, the 6 French samples of the first 42 writers are used in training while the 6 Arabic samples of the other 42 writers are used in testing. In a similar fashion, the last experiment

Table 6 Classification rates of script-dependent evaluations on the QUWI and MSHD databases

Training data	Test data	Features	Classification rate	
			SVM (%)	ANN (%)
QUWI-English	QUWI-English	Slant and Curv.	68.50	66.50
		Texture	60.00	61.50
		Fractal	63.50	63.00
QUWI-Arabic	QUWI-Arabic	Slant and Curv.	68.50	65.00
		Texture	61.50	62.50
		Fractal	61.50	61.50
MSHD-French	MSHD-French	Slant and Curv.	67.06	69.44
		Texture	70.63	68.25
		Fractal	62.30	61.51
MSHD-Arabic	MSHD-Arabic	Slant and Curv.	76.98	73.41
		Texture	70.63	71.41
		Fractal	61.51	62.30

Table 7 Classification rates of script-independent evaluations on the QUWI and MSHD databases

Training data	Test data	Features	Classification rate	
			SVM (%)	ANN (%)
QUWI-English	QUWI-Arabic	Slant and Curv.	60.00	64.00
		Texture	60.00	62.50
		Fractal	67.00	65.00
QUWI-Arabic	QUWI-English	Slant and Curv.	60.50	65.00
		Texture	54.00	60.00
		Fractal	62.50	63.00
MSHD-French	MSHD-Arabic	Slant and Curv.	69.05	69.44
		Texture	69.05	68.65
		Fractal	70.63	69.84
MSHD-Arabic	MSHD-French	Slant and Curv.	57.14	61.90
		Texture	52.38	60.71
		Fractal	57.94	61.11

involves 6 Arabic writings of first 42 writers in the training and the 6 French writings of remaining 42 writers in the test data set.

The results of script-dependent and script-independent experiments are summarized in Tables 6 and 7, respectively. Comparing the classification rates across Tables 6 and 7, it can be seen that the script-dependent evaluations give better performances than the script-independent evaluations. Naturally, when writing samples in same script are used for training and test, the system achieves better classification rates as compared to those obtained when using different scripts for training and test. Similar to text-independent vs. text-dependent evaluations, the orientation and curvature features perform better than texture and fractal features but are also more sensitive to the script under study.

6.3 Cross-database evaluations

In the final series of experiments, we use the writing samples of one database in training and those of the other database in testing. Since the two databases have been developed in two different countries, it would be very interesting to analyze if male/female writers from geographically different regions share some common characteristics.

We first use all Arabic samples in the QUWI database as training set and all Arabic samples in the MSHD database as the test set. For the second experiment, the training and test data sets are reversed. Since English and French share the same script with minor variations, in the next series of experiments we use all English samples of QUWI database for training and all French samples of the MSHD database for testing. Similarly, for completeness we also reverse the scenario using all French samples of MSHD database as training and all English samples of QUWI database as test set. This distribution ensures that these cross-database evaluations compare different textual contents in the same script and the classification performance is not affected by the script variations in training and test sets. The results of these evaluations are summarized in Table 8.

An inspection of results in Table 8 reveals that acceptable classification rates are achieved when the system is trained on writing samples in one database and tested on a totally different database. The classification rates are better when QUWI datasets (Arabic/English) are used for training and MSHD datasets (French/Arabic) are used for testing as compared to the reverse case (MSDH datasets for training and QUWI datasets for testing). This can be attributed to the fact that size of test dataset is approximately twice the size of training dataset in the later case hence resulting in relatively low classification rates. Considering the fact that the two databases have been produced by writers in totally different geographical locations and cultural circumstances, classification rates of as high as 72 % (on Arabic) and 63.5 % on (English-French) are very encouraging. They are also indicative of the fact that writers belonging to a particular gender (male/female) do share some common characteristics which are, to some extent, consistent across individuals from different backgrounds.

Table 8 Classification rates of cross-database evaluations

Training data	Test data	Features	Classification rate	
			SVM (%)	ANN (%)
QUWI-Arabic	MSHD-Arabic	Slant and Curv.	72.22	70.04
		Texture	68.85	69.44
		Fractal	57.14	60.00
MSHD-Arabic	QUWI-Arabic	Slant and Curv.	55.13	58.88
		Texture	58.13	58.37
		Fractal	53.13	56.25
QUWI-English	MSHD-French	Slant and Curv.	57.74	60.52
		Texture	61.51	63.49
		Fractal	62.90	62.50
MSHD-French	QUWI-English	Slant and Curv.	56.75	58.13
		Texture	57.87	57.87
		Fractal	54.25	54.75

Table 9 Performance comparison of gender prediction methods

Study	Database	Training data	Test data	Results (%)
Bandi and Srihari [2]	CEDAR letter	800	400	77.5
Hamid and Loewenthal [18]	English and Urdu texts	–	30	68
Liwicki et al. [28]	IAM-onDB (Offline)	80	50	55.39
	IAM-onDB (Offline + Online)			67.57
Proposed method	QUWI	300	100	68.75
	MSHD	42	42	73.02

We also present a comparative analysis of the proposed system with the existing systems on this problem. A quantitative comparison of the classification rates achieved by different systems is summarized in Table 9. While Bandi and Srihari [2] report a classification rate of 77.5 % on a large database, it should be noted that these results are based on the CEDAR letter where each individual copied the same fixed text thrice to constitute the database. In addition, some of the features used in this study are based on comparing text of known semantic content [47]. The results of [18] are based on examination by human judges and are not automated. The only meaningful comparison, therefore, can be made with the study in [28] where the authors evaluate the classification performance on 50 individuals in the test set and 80 in the training set. A classification rate of 67.57 % is reported when combining the online and offline features while it drops to 55.39 % when using only offline features. Our proposed system not only realizes better performances than Liwicki et al.'s [28] but has also been evaluated on a larger dataset.

In addition to the classification rates, another interesting and novel aspect of our study is the analysis of classification performance in a number of interesting scenarios including text-dependent, text-independent, script-dependent, script-independent and cross-database evaluations. To the best of authors' knowledge, this is the first study of its kind to consider such experimental scenarios for gender classification task.

In summary, the series of experiments that we conducted validate the hypothesis that correlation does exist between handwriting and the gender of its writer. Among the three categories of features that we employed, orientation and curvature features proved to be the most effective in a number of evaluation scenarios. The text-dependent vs. text-independent and script-dependent vs. script-independent evaluations revealed the effectiveness of the proposed features in predicting gender in the aforementioned scenarios. Finally, the classification rates of the cross-database evaluations reflect that gender can be predicted with acceptable success rates independent of the background of the writer in question. A comparison with existing methods on this subject also reveals the effectiveness of our system.

7 Conclusion

This study presented an effective method for gender classification from handwriting. Although a popular research area in psychological studies for many decades, this problem is relatively less explored by researchers in computer sciences. We identify a subset of discriminative writing attributes suggested in different psychological studies and algorithmically compute features from scanned images of writing samples to estimate these attributes. The three categories of features that we consider in our study include orientation and curvature features, texture-based features and the fractal dimensions. These features were used to train two classifiers: ANN and SVM. The effectiveness of these features in predicting the gender of the writer of a given sample was evaluated on two databases: the QUWI and the MSHD. The performance of each type of features was analyzed separately to study its usefulness in predicting the gender of the writer of a writing in question. The use of two different databases with text samples in different languages also allowed studying the sensitivity of the classification performance in different evaluation scenarios. These included considering the same/different textual content in training and test images and having writing samples in same/different scripts for training and test sets. In addition, an analysis of the cross-database evaluations was also carried out revealing that the writings of the two gender groups (male and female) share some common attributes which are consistent across individuals from different backgrounds. The average classification rates on these distinct evaluation scenarios are very encouraging and support the arguments put forward in this work.

In our further study on the subject, we intend to study the prediction of other attributes of writers from their handwritings, especially handedness (left or right) and age. It would also be interesting to introduce additional features and then apply a feature selection mechanism to determine which are the most discriminative features for this and other similar problems. A combination of different classifiers to enhance the overall classification rates may also be explored.

References

1. Al Ma'adeed S, Ayouby W, Hassaine A, Aljaam JM (2012) Quwi: An Arabic and English handwriting dataset for offline writer identification. In: Proceedings of 13th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp 746–751
2. Bandi KR, Srihari SN (2005) Writer demographic classification using bagging and boosting. In: Proceedings 12th International Graphonomics Society Conference, pp 133–137
3. 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. Personal Individ Differ 39(2):459–468
4. Bertolini D, Oliveira LS, Justino E, Sabourin R (2013) Texture-based descriptors for writer identification and verification. Expert Systems Appl 40(6):2069–2080
5. Blumenstein M, Liu XY, Verma B (2007) An investigation of the modified direction feature for cursive character recognition. Pattern Recognit 40(2):376–388
6. Blumenstein M, Verma B, Basli H (2003) A novel feature extraction technique for the recognition of segmented handwritten characters. In: Proceedings of the Seventh International Conference on Document Analysis and Recognition, pp 137–141
7. BoulTtreau V, Vincent N, Sabourin R, Emptoz H (1998) Handwriting and signature: One or two personality identifiers? In ICPR '98: Proceedings of the 14th International Conference on Pattern Recognition, pp 1758–1760
8. Broom ME, Thompson B, Bouton MT (1929) Sex differences in handwriting. J Appl Psychol 13(2):159–166
9. Burr V (2002) Judging gender from samples of adult handwriting: accuracy and use of cues. J Soc Psychol 142(6):691–700
10. Canu S, Grandvalet Y, Guigue V, Rakotomamonjy A (2005) Svm and kernel methods matlab toolbox. Perception Systemes et Information, INSA de Rouen, Rouen
11. Cha S-H, Srihari SN (2001) A priori algorithm for sub-category classification analysis of handwriting. In: Proceedings of the 6th International Conference on Document Analysis and Recognition, pp 1022–1025
12. Chaabouni A, Boubaker H, Kherallah M, Alimi AM, El Abed H (2010) Fractal and multi-fractal for Arabic offline writer identification. In: 20th International Conference on Pattern Recognition, pp 3793–3796
13. Deguchi K (1986) Two-dimensional auto-regressive model for analysis and sythesis of gray-level textures. In: Proceedings of the 1st International Symposium for Science on Form, pp 441–449
14. Djeddi C, Siddiqi I, Souici-Meslati L, Ennaji A (2013) Codebook for writer characterization: A vocabulary of patterns or a mere representation space? In: Proceedings of the 12th Int'l Conference on Document Analysis and Recognition
15. Furnham Adrian, Gunter Barrie (1987) Graphology and personality: another failure to validate graphological analysis. Personal Indivi Differ 8(3):433–435
16. Garain U, Paquet T (2009) Off-line multi-script writer identification using ar coefficients. In: 10th International Conference on Document Analysis and Recognition, pp 991–995
17. Goodenough FL (1945) Sex differences in judging the sex of handwriting. J Soc Psychol 22:61–68
18. Hamid S, Loewenthal KM (1996) Inferring gender from handwriting in Urdu and English. J Soc Psychol 136(6):778–782
19. Hanusiak RK, Oliveira LS, Justino E, Sabourin R (2012) Writer verification using texture-based features. IJDAR 15(3):213–226
20. Hartley James (1991) Sex differences in handwriting: a comment on spear. Br Educ Research J 17(2):141–145
21. Hayes William N (1996) Identifying sex from handwriting. Percept Motor Skills 83:91–800
22. Huang D, Shan C, Ardabilian M, Wang Y, Chen L (2011) Local binary patterns and its application to facial image analysis: a survey. IEEE Trans Systems Man Cybern Part C Appl Reviews 41(6):765–781
23. Huber RA, Headrick AM (1999) Handwriting identification: facts and fundamentals. CRC Press, Boca Raton
24. Kimura F, Kayahara N, Miyake Y, Shridhar M (1997) Machine and human recognition of segmented characters from handwritten words. In: Proceedings of the 4th International Conference on Document Analysis and Recognition, pp 866–869
25. King RN, Koehler DJ (2000) Illusory correlations in graphological inference. J Exp Psychol Appl 6(4):336–348
26. Klimoski RJ, Rafaeli A (1983) Inferring personal qualities through handwriting analysis. J Occup Psychol 56(3):191–202
27. Lin TY, Xie Y, Wasilewska A, Liau CJ (2008) Data mining: foundations and practice, vol 118. Springer
28. Liwicki M, Schlapbach A, Bunke H (2011) Automatic gender detection using on-line and off-line information. Pattern Anal Appl 14(1):87–92
29. Liwicki M, Schlapbach A, Bunke H, Bengio S, Marièthoz J, Richiardi J (2006) Writer identification for smart meeting room systems. In: Proceedings of 7th IAPR Workshop on Document Analysis Systems, vol 3872, pp 186–195
30. Lopes R, Betrouni N (2009) Fractal and multifractal analysis: a review. Med Image Anal 13(4):634–649
31. Dehkordi ME, Sherkat N, Allen T (2003) Handwriting style classification. Int J Document Anal Recognit 6:55–74
32. Mhidra H, Brochard J, Leard M (1993) Ar models and bidimensional discrete moments applied to texture modelling and recognition. Pattern Recognit 26(5):721–726
33. Yamada H, Nakano Y (1996) Cursive handwritten word recognition using multiple segmentation determined by contour analysis. IEICE Trans Inf Systems E79-D(5):464–470
34. Neter Efrat, Ben-Shakhar Gershon (1989) The predictive validity of graphological inferences: a meta-analytic approach. Personal Individ Differ 10(7):737–745
35. Newhall SM (1926) Sex differences in handwriting. J Appl Psychol 10(2):151–161

36. Oe Shunichiro (1993) Texture segmentation method by using two-dimensional AR model and kullback information. Pattern Recognit 26(2):237–244
37. Ojala T, Pietikainen M, Harwood D (1994) Performance evaluation of texture measures with classification based on kullback discrimination of distributions. In: Proceedings of the 12th IAPR International Conference on Pattern Recognition, vol 1, pp 582–585
38. Ojala T, Pietikainen M, Harwood D (1996) A comparative study of texture measures with classification based on feature distributions. Pattern Recognit 29:51–59
39. Ojala T, Pietikainen M, Maenpaa T (2002) Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans Pattern Anal Mach Intell 24(7):971–987
40. Sabourin R, Emptoz H, Vincent N, Bouletreau V (2000) How to use fractal dimensions to qualify writings and writers. Fractals 08(01):85–97
41. Said HES, Tan TN, Baker KD (2000) Personal identification based on handwriting. Pattern Recognit 33:149–160
42. Shackleton V, Newel S (1994) European management selection methods: a comparison of five countries. Int J Sel Assess 2(2):91–102
43. Siddiqi I, Vincent N (2009) A set of chain code based features for writer recognition. In: Proceedings of 10th International Conference on Document Analysis and Recognition, pp 981–985
44. Siddiqi I, Vincent N (2010) Text independent writer recognition using redundant writing patterns with contour-based orientation and curvature features. Pattern Recognit 43(11):3853–3865
45. Siddiqi I, Vincent N (2008) Combining global and local features for writer identification. In: Proceedings of the Eleventh International Conference on Frontiers in Handwriting Recognition, pp 48–53
46. Sokic E, Salihbegovic A, Ahic-Djokic M (2012) Analysis of off-line handwritten text samples of different gender using shape descriptors. In: Proceedings of IX International Symposium on Telecommunications (BIHTEL), pp 1–6
47. Srihari SN, Cha S-H, Arora H, Lee S (2002) Individuality of handwriting. J Forensic Sci 47(4)
48. Tenwolde H (1934) More on sex differences in handwriting. J Appl Psychol 18:705–710
49. Tett RP, Palmer CA (1997) The validity of handwriting elements in relation to self-report personality trait measures. Personal Individ Differ 22(1):11–18
50. Vicsek T (1990) Mass multifractals. Physica A Stat Mech Appl 168(1):490–497
51. Vincent N, Emptoz H (1995) Fractal reviews in the natural and applied sciences, chap. A classification of writings based on fractals, Chapman and Hall, pp 320–331
52. Wall Karin, Danielsson Per-Erik (1984) A fast sequential method for polygonal approximation of digitized curves. Comput Vision Graphics Image Process 28(3):220–227
53. Wang X, Han TX, Yan S (2009) An hog-lbp human detector with partial occlusion handling. In: Proceedings of IEEE 12th International Conference on Computer Vision, pp 32–39