EURASIP Journal on Image and Video Processing
a SpringerOpen Journal

RESEARCH Open Access

Automatic prediction of age, gender, and nationality in offline handwriting

Somaya Al Maadeed* and Abdelaali Hassaine

Abstract

The classification of handwriting into different categories, such as age, gender, and nationality, has several applications. In forensics, handwriting classification helps investigators focus on a certain category of writers. However, only a few studies have been carried out in this field. Classification of handwriting into a demographic category is generally performed in two steps: feature extraction and classification. The performance of a system depends mainly on the feature extraction step because characterizing features makes it possible to distinguish between writers. In this study, we propose several geometric features to characterize handwritings and use these features to perform the classification of handwritings with regards to age, gender, and nationality. Features are combined using random forests and kernel discriminant analysis. Classification rates are reported on the QUWI dataset, reaching 74.05% for gender prediction, 55.76% for age range prediction, and 53.66% for nationality prediction when all writers produce the same handwritten text and 73.59% for gender prediction, 60.62% for age range prediction, and 47.98% for nationality prediction when each writer produces different handwritten text.

Keywords: Writer demographic category classification; Handwriting analysis; Chain code; Edge-based directional features; Writer identification

1 Introduction

Handwritings can be classified into many categories including gender, age, handedness, and nationality. This type of classification has several applications. For example, in the forensic domain, handwriting classification can help the investigators to focus on a certain category of suspects. Additionally, processing each category separately leads to improved results in writer identification and verification applications.

There are only a few studies in the literature that investigate the automatic detection of gender, age, and handedness from handwritings. Bandi et al. [1] proposed a system that classifies the handwritings into demographic categories using the 'macro-features' introduced in [2]. These features focus on measures such as pen pressure, writing movement, stroke formation, and word proportion. The authors reported classification accuracies of 77.5%, 86.6%, and 74.4% for gender, age, and handedness classification, respectively. However, in this study, all the writers had to produce the same letter. Unfortunately, this is not always the case in real forensic caseworks. Moreover, the dataset used in this study is not publicly available.

Liwicki et al. [3] also performed the classification of gender and handedness in the online mode (which means that the temporal information about the handwriting is available). The authors used a set of 29 features extracted from the online information and its offline representation and applied support vector machines and Gaussian mixture models to perform the classification. The authors reported a performance of 67.06% for gender classification and 84.66% for handedness classification. In a recent study [4], the authors reported separately the performance of the offline mode, the online mode and their combination. The performance reported for the offline mode was 55.39%, which is slightly better than chance.

In this paper, we propose a new method for the detection of the age range, gender, and nationality of the writer of a handwritten document. A set of novel features are proposed and described, including directions, curvatures, tortuosities, chain codes, and edge-based directional features. These features are combined using several classifiers, including random forests and kernel discriminant

* Correspondence: s_alali@qu.edu.qa
Computer Science and Engineering Department, College of Engineering, Qatar University, Doha, Qatar

analysis. This method is evaluated using the QUWI database, which is the only available public dataset containing annotations regarding gender, age range, and nationality.

The remainder of this paper is organized as follows: Sections 2 and 3 give a detailed description of our feature extraction and classification methods. Section 4 presents the dataset used in this study and the detailed results. Section 5 concludes this work and draws some perspectives. Our method consists of two main steps: feature extraction and classification. These two steps are illustrated in Figure 1.

2 Feature extraction

In this step, the characterizing features are extracted from the handwriting. To make the system pen independent, images are first binarized using the Otsu thresholding algorithm [5]. The following subsections describe the features considered in this study. These features do not correspond to a single value, but are defined by a probability distribution function (PDF) extracted from the handwriting images to characterize the writer's individuality [6,7]. The PDF describes the relative likelihood for a certain feature to take on a given value.

Note that all these developed features or their equivalents are used by forensic document examiners as well as graphologists in order to distinguish between different categories of writers [8].

1.1 Direction feature (f1)

This method has been used in writer identification [7,9], and its implementation closely resembles the one proposed by Matas et al. [10]. First, we compute the Zhang skeleton of the binarized image. This skeleton is well known for not producing parasitic branches unlike most skeletonization algorithms [11]. The skeleton is then segmented at its junction pixels. Then, we traverse the pixels of the obtained segments of the skeleton using the predefined order favoring the four-connectivity neighbors as shown in Figure 2a. A result of such an ordering is shown in Figure 2b. For each pixel p, we consider the $2 \cdot N + 1$ neighboring pixels centered at position p. A linear regression of these pixels gives a good estimation of the tangent at the pixel p (Figure 2c). The value of N has empirically been set to 5 pixels throughout this paper.

The PDF of the resulting directions is computed as a vector of probabilities for which the size has been empirically set to 10. It is worth noting that this is the first time that such a method of computing directions has been proposed for categorization applications.

2.2 Curvature feature (f2)

In forensic document examination, curvature is commonly accepted as a characterizing feature [7,8]. We have adapted this method to handwriting as follows: for each pixel p belonging to the contour, we consider a neighboring window of size t. Inside this window, we compute the number of pixels n_1 and the number of pixels n_2 that belong to the background and foreground, respectively (see Figure 3a). The difference $n_1 - n_2$ is positive at the points on which the contour is convex and negative at the points on which the contour is concave and is therefore a good indicator of the local curvature of the contour. Therefore, we estimate the curvature as being: $C = \frac{n_1 - n_2}{n_1 + n_2}$. The value C is illustrated in Figure 3b on a binary shape for which t has been empirically set to 5. The PDF of curvatures is computed in a vector with a size empirically set to 100. This way of computing curvatures is also novel in the field of offline writer identification and categorization, and to the extent of our knowledge, it has never been used before.

2.3 Tortuosity feature (f3)

This feature makes it possible to distinguish between fast writers who produce smooth handwriting and slow writers who produce ‘tortuous’/twisted handwriting. To estimate tortuosity, for each pixel p of the text, we determine the longest line segment that traverses p and is completely included inside the foreground (Figure 4a). An example of estimated tortuosities is shown in Figure 4b.

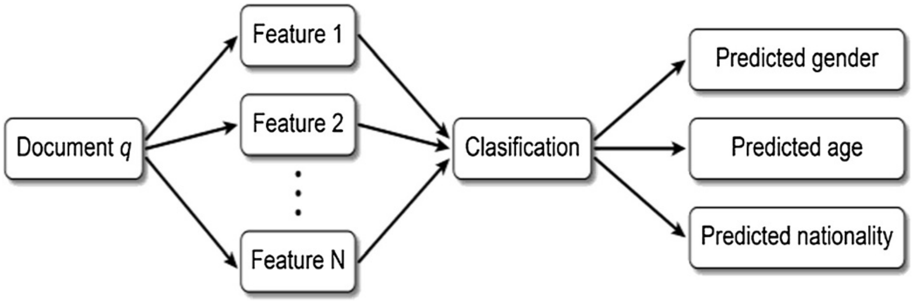

Figure 1 General scheme of our method.

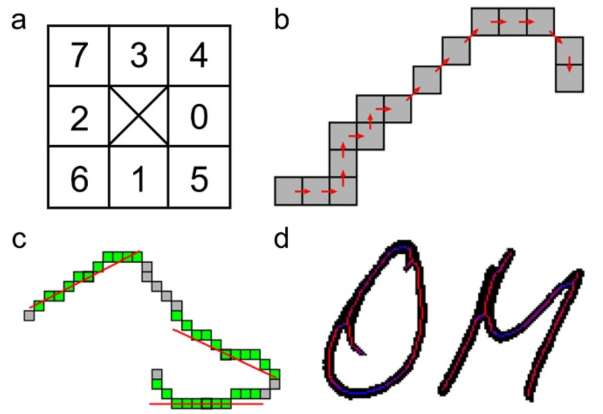

Figure 2 Computing local directions. The predefined order for traversing shapes **(a)**. Example of an ordered shape **(b)**. Estimating directions by computing linear regression of neighbors of the three pixels in bold **(c)**. Binary image and its corresponding Zhang skeleton **(d)**; the red color corresponds to a π/2 tangent, and the blue color corresponds to a zero tangent.

The PDF of the angles of the longest traversing segments is produced in a vector with the size set to 10, as mentioned previously.

2.4 Chain code features (f4 to f7)

Chain codes are generated by scanning the contour of the text and assigning a number to each pixel according to its location with respect to the previous pixel. Figure 5 shows a contour and its corresponding chain code.

Chain codes have been applied to writer identification in [12]. These features make it possible to characterize a detailed distribution of curvatures in the handwriting. Chain codes might be applied at different orders:

f4: The PDF of i patterns in the chain code list such that $i \in 0,1,\ldots,7$. This PDF has a size of 8.
f5: The PDF of (i, j) patterns in the chain code list such that $i,j \in 0,1,\ldots,7$. This PDF has a size of 64.

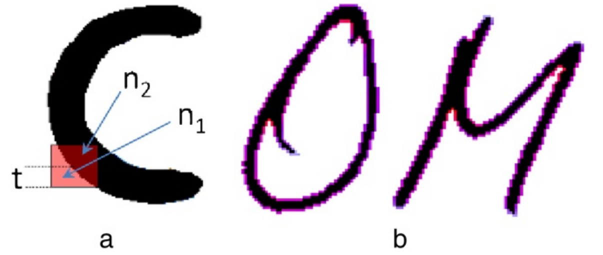

Figure 3 Computing curvatures (a) and curvatures highlighted on binary image (b). Red corresponds to the maximum curvature and blue corresponds to the minimum one.

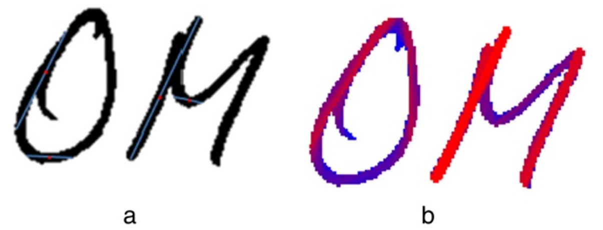

Figure 4 Computing tortuosity. Longest traversing segment for four different pixels **(a)**. Length of maximum traversing segment: red corresponds to the maximum length, blue to the minimum one **(b)**.

Similarly, f6 and f7 correspond to the PDF of (i, j, k) and (i, j, k, l) in the chain code list with sizes of 512 and 4,096, respectively. Not all successions of chain code patterns can be obtained. For example, the chain code pattern (1, 5) is not a possible succession, and therefore its corresponding distribution in the PDF will always be nil.

2.5 Edge-based directional features (f8 to f26)

Initially introduced in [9], these features provide a detailed distribution of directions and can also be applied at several sizes by positioning a window centered at each contour pixel and counting the occurrences of each direction, as shown in Figure 6a. These features have been computed from size 1 (f8, which has a PDF size of 4) to size 10 (f17, which has a PDF size of 40). We have also extended these features to include not only the contour of the moving window but also the whole window (Figure 6b) [7]. This feature has been computed from size 2 (f18, which has a PDF size of 12) to size 10 (f26, which has a PDF size of 220).

3 Classification

In this step, the features previously presented are used to decide which category each handwriting belongs to. When performing the classification, each element of the feature vectors will be used as a separate input for the classifier. (For example, f1 will be an input vector of 10 elements for the classifier.)

We have combined these features using a Random Forest classifier [13] with kernel discriminant analysis using spectral regression (SR-KDA). Descriptions of the random forests classifier and the SR-KDA [14] are given below.

The use of these two classifiers is justified by their ability to train on large datasets for features and achieving high classification rates [15].

3.1 Random forest classifier

Random forests is an ensemble learning method for classification that operates by constructing a multitude of decision trees at training time and outputting the class that is

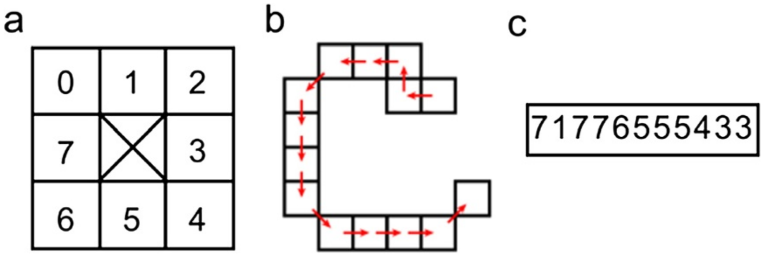

Figure 5 Order followed to generate chain code (a), example shape (b), and its corresponding chain code (c).

the mode of the classes output by individual trees. Each decision tree is constructed as follows:

1. If the number of cases in the training set is N, sample n cases such as $n < N$ at random but with replacement from the original data. This sample will be the training set for growing the tree.
2. If there are M input variables, a number $m < < M$ is specified such that at each node, m variables are selected at random from M and the best split on these m is used to split the node. The value of m is held constant during the forest growing.
3. Each tree is grown to the largest extent possible. There is no pruning.

In our case, we built the random forest classifiers for the cases of age, gender, and nationality using the *R* random forest library [16].

3.2 Kernel discriminant analysis using spectral regression

Let $x_i \in R^d$, $i = 1,\ldots, m$ be training vectors represented as an $m \times m$ kernel matrix K such that $\mathrm{K}(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle$, where $\Phi(x_i)$ and $\Phi(x_j)$ are the embeddings of data items x_i and x_j. If v denotes a projective function into the kernel feature space, then the objective function for KDA is [17]:

$$\max_v D(v) = \frac{v^T C_b v}{v^T C_t v}, \qquad (1)$$

where C_b and C_t denote the between-class and total scatter matrices in the feature space, respectively. Equation 1 can be solved by the eigen-problem $C_b = \lambda C_t$. It is proved in [18] that Equation 1 is equivalent to:

$$\max_\alpha D(\alpha) = \frac{\alpha^T \mathrm{KWK}_\alpha}{\alpha^T \mathrm{KK}_\alpha}, \qquad (2)$$

where $\alpha = [\alpha_1, \alpha_2,\ldots, \alpha_m]^T$ is the eigenvector satisfying $\mathrm{KWK}\alpha = \lambda \mathrm{KK}\alpha$.

$\mathrm{W} = (\mathrm{W}_l)_{l = 1,\ldots,n}$ is an $(m \times m)$ block diagonal matrix of labels arranged such that the upper block corresponds to positive examples and the lower one corresponds to negative examples of the class. Each eigenvector α yields a projection function v in the feature space.

It is also shown in [4] that instead of solving the eigenproblem in KDA, the KDA projections can be obtained by the following two linear equations:

$$\mathrm{W}\phi \;=\; \lambda\phi \;\; (\mathrm{K} + \delta \mathrm{I})\alpha \;=\; \phi \qquad (3)$$

where ϕ is an eigenvector of W, I is the identity matrix, and $\delta > 0$ is a regularization parameter. $\mathrm{W} = (\mathrm{W}_l)_{l = 1,\ldots,n}$ is an $(m \times m)$ block diagonal matrix of labels arranged such that the upper block corresponds to positive examples and the lower one corresponds to negative examples of the class. Eigenvectors ϕ are obtained directly from the Gram-Schmidt method. Because $(\mathrm{K} + \delta \mathrm{I})$ is positive definite, a Cholesky decomposition is used to solve the

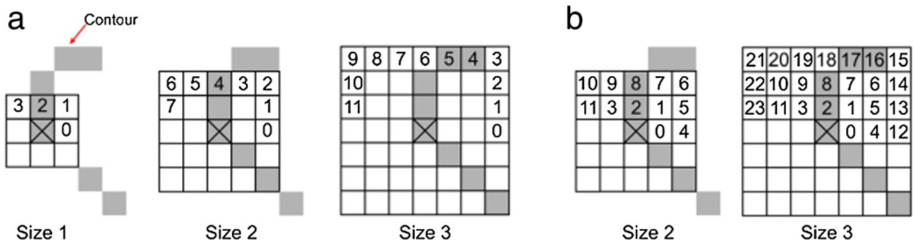

Figure 6 Counting edge-based directional features when considering contour of the moving window (a) and whole moving window (b).

linear equations in (3). Thus, for the resolution of the linear system of Equation 3, the system becomes:

$$K+\delta \mathrm{I})\alpha=\phi \Leftrightarrow \begin{cases} R^T\theta=\phi \\ R\alpha=\theta \end{cases} \tag{4}$$

i.e., solve the system to first find vector θ and then find vector α. In summary, SR-KDA only needs to solve a set of regularized regression problems, and there is no eigenvector computation involved. This results in a significant improvement of computational complexity and allows large kernel matrices to be handled. After obtaining α, the decision function for the new data item is calculated from:

$$f(x)=\sum_{i=1}^{n}\alpha_i \mathrm{K}(x,x_i). \tag{5}$$

The classification results of those classifiers for all the presented features on the QUWI dataset will be shown in the next Section.

Table 1 Correct classification rates for gender detection using random forests (RF) and kernel discriminant analysis (KDA)

	Arabic				English				Both			
	Same text (%)		Different text (%)		Same text (%)		Different text (%)		Same text (%)		Different text (%)	
Feature	RF	KDA	RF	KDA	RF	KDA	RF	KDA	RF	KDA	RF	KDA
f1	59.4	65.5	62.1	62.9	62.0	64.1	63.3	62.9	63.3	64.7	63.0	59.2
f2	68.3	63.9	66.9	63.1	63.4	62.8	69.2	62.2	64.9	68.5	67.5	60.6
f3	58.1	59.9	62.1	60.6	62.7	58.3	63.0	54.3	61.3	60.2	61.1	59.9
f4	53.0	60.6	54.8	57.1	64.1	53.5	59.2	53.3	57.8	61.6	56.1	62.7
f5	68.9	64.8	65.2	65.3	65.5	62.4	66.8	65.7	66.6	69.2	67.0	64.4
f6	67.0	68.6	68.6	66.1	67.3	69.3	71.6	70.2	66.9	68.2	70.3	68.0
f7	65.7	70.0	68.3	71.8	68.7	69.3	69.9	73.7	66.3	74.1	69.3	73.6
f8	57.1	58.9	54.8	59.1	52.8	57.2	58.5	54.9	54.6	58.5	56.3	60.2
f9	55.2	58.2	56.6	57.4	59.9	57.2	60.6	54.9	57.8	57.8	57.5	59.9
f10	55.9	58.7	56.6	61.9	57.8	57.2	55.7	60.0	56.3	58.8	54.4	59.5
f11	58.4	61.5	59.7	58.1	59.9	61.7	56.8	60.6	57.8	61.6	58.7	56.0
f12	63.8	61.5	62.8	62.4	64.1	58.6	59.5	62.9	62.6	63.0	60.8	59.5
f13	64.4	62.0	64.5	62.8	63.7	63.5	63.0	60.3	63.9	62.3	63.0	60.9
f14	63.8	63.6	65.2	60.8	64.4	60.0	62.3	61.3	63.3	63.3	64.9	62.0
f15	64.4	63.6	66.9	62.9	65.1	66.2	61.6	62.5	65.3	59.2	66.2	67.3
f16	65.1	63.0	69.3	63.8	62.7	63.8	64.7	61.3	63.6	60.6	67.7	62.3
f17	66.7	61.8	68.6	59.4	61.6	66.2	64.4	61.3	62.6	57.8	67.4	60.2
f18	56.5	59.1	58.6	59.6	61.3	58.3	60.9	56.8	58.1	59.2	63.2	62.7
f19	56.5	61.8	58.3	62.3	60.6	55.5	59.5	61.9	57.9	65.1	60.3	61.6
f20	60.3	64.1	60.3	63.8	58.8	59.7	60.9	64.8	60.1	67.5	60.3	60.6
f21	62.5	64.8	62.4	64.1	62.7	62.4	61.6	63.2	61.3	67.5	63.0	62.3
f22	65.7	65.3	63.5	63.6	62.7	61.7	61.9	63.2	64.4	65.4	63.6	64.1
f23	67.9	63.7	63.5	66.1	67.6	62.1	63.0	65.4	66.4	63.3	64.1	64.1
f24	67.9	64.4	66.2	65.9	70.8	63.1	64.4	64.1	68.8	63.7	65.5	65.9
f25	68.3	66.5	65.2	66.8	69.7	64.8	64.7	65.4	68.3	66.1	66.3	65.1
f26	68.3	66.8	67.6	66.4	69.7	66.9	65.7	68.6	69.3	64.7	66.7	64.1
f1 + f2 + f3	65.4	68.6	66.6	69.5	64.8	66.6	74.7	68.3	62.8	72.0	69.1	64.1
f4 +,…, + f7	67.3	68.9	68.6	70.1	69.0	69.0	70.2	70.8	64.8	70.2	69.8	69.0
f8 +,…, + f17	67.6	65.8	67.6	63.6	66.2	64.5	63.7	63.8	66.9	63.7	67.7	66.9
f18 +,…, + f26	67.3	67.2	64.8	66.8	71.1	64.5	64.0	66.4	68.8	66.8	66.1	65.1
f1 +,…, + f26	71.1	68.4	69.0	71.6	69.7	68.6	68.2	66.4	69.8	72.3	68.7	70.8

4 Evaluation

In this section, we describe the QUWI handwriting database on which the experiments have been conducted. We also present the results obtained for each individual feature as well as their combination using random forests and kernel discriminant analysis. The results are then analyzed and discussed.

4.1 Dataset

To the best of our knowledge, the only publicly available handwriting dataset annotated with respect to age, gender, and nationality is the QUWI dataset [19]. This dataset contains handwritings of 1,017 writers in both English and Arabic. In each language, writers produced one text that is the same for all the writers and another text that is different for every writer. Moreover, writers in this dataset have different genders, age ranges, and nationalities. Because very few writers are left-handed (around fifty writers), this dataset can only be useful for handedness detection.

To perform the classification, 70% of this dataset has been used for training and 30% for testing as is often the case in data mining [18]. We have computed the presented

Table 2 Correct classification rates for age range detection using random forests (Rf) and kernel discriminant analysis (Kda)

	Arabic				English				Both			
	Same text (%)		Different text (%)		Same text (%)		Different text (%)		Same text (%)		Different text (%)	
Feature	RF	KDA	RF	KDA	RF	KDA	RF	KDA	RF	KDA	RF	KDA
f1	53.7	56.5	56.9	52.1	47.2	58.3	53.6	49.2	50.4	52.9	53.2	47.2
f2	56.2	52.0	59.3	53.6	51.1	51.4	57.4	53.0	55.4	49.8	58.7	50.4
f3	55.9	54.9	56.2	49.9	49.7	52.1	53.3	51.1	52.4	54.0	53.5	45.4
f4	56.5	54.4	50.7	51.9	49.3	53.1	50.9	51.1	51.9	50.5	52.5	46.8
f5	55.9	48.9	56.6	48.9	50.4	43.8	54.0	48.9	54.1	43.6	55.8	47.2
f6	58.7	48.2	60.7	48.6	50.4	49.3	55.4	50.5	55.1	45.7	58.4	46.1
f7	59.1	53.9	61.4	55.3	50.7	53.1	56.8	54.6	54.6	51.6	58.7	48.9
f8	49.8	57.2	51.7	52.6	48.2	55.9	50.5	56.5	49.6	55.4	52.0	49.7
f9	56.2	53.9	55.5	53.4	49.7	56.9	54.7	56.2	54.8	48.8	54.6	47.2
f10	57.1	56.0	55.5	52.9	48.2	53.8	52.6	54.0	54.4	55.7	54.8	48.6
f11	55.6	53.4	55.2	54.8	47.9	47.2	52.9	56.5	53.4	54.3	53.2	49.3
f12	54.0	53.5	53.1	53.3	50.4	49.7	56.4	53.7	52.8	52.9	54.9	52.5
f13	55.2	53.5	55.5	52.6	50.4	47.9	55.7	53.0	52.6	54.0	56.1	47.9
f14	56.2	53.7	54.1	52.4	53.2	51.0	54.7	55.9	54.1	51.2	56.5	47.9
f15	56.2	50.8	56.2	52.3	52.8	48.6	54.7	50.5	54.4	51.9	57.2	51.1
f16	55.9	49.6	57.2	52.8	50.4	48.3	57.8	48.6	53.9	50.9	56.3	51.4
f17	55.2	51.0	58.3	52.8	48.9	49.0	55.4	50.2	53.4	50.9	54.9	50.4
f18	55.6	54.1	58.6	52.8	49.3	55.2	55.1	56.2	53.1	49.8	55.1	49.3
f19	57.1	54.9	56.6	53.8	50.4	53.8	55.4	56.2	54.4	51.9	55.3	48.6
f20	58.1	53.0	55.2	52.9	50.7	51.0	54.2	54.9	54.4	52.6	54.6	49.7
f21	57.5	51.3	54.1	53.4	49.7	48.3	54.6	56.8	54.3	51.6	54.2	50.0
f22	57.1	50.3	55.9	53.8	50.7	47.2	53.7	54.3	54.3	50.9	54.1	50.4
f23	55.2	49.9	58.3	53.3	49.7	47.6	54.9	53.3	53.1	51.9	54.8	47.5
f24	54.9	50.3	58.3	52.4	50.7	48.3	56.5	53.0	53.8	51.9	56.3	49.7
f25	56.2	50.4	60.0	54.4	51.8	49.3	58.0	53.3	54.1	53.6	58.6	51.1
f26	56.8	51.8	58.6	54.8	51.8	49.3	58.7	51.4	54.9	53.6	57.7	51.4
f1 + f2 + f3	58.7	53.9	62.4	50.4	52.8	54.1	59.9	49.5	55.8	54.0	60.6	46.8
f4 +,…, + f7	58.4	50.8	60.7	52.3	50.0	48.6	56.4	53.0	54.4	47.1	58.5	48.6
f8 +,…, + f17	57.1	49.1	58.3	54.1	48.9	50.7	54.7	51.4	54.1	49.5	56.7	53.2
f18 +,…, + f26	55.9	50.1	57.2	53.9	50.7	49.7	56.4	55.9	53.1	53.3	56.8	49.3
f1 +,…, + f26	58.1	53.0	59.3	55.8	51.4	53.5	57.4	53.3	55.1	49.8	59.4	53.9

features on this dataset. As mentioned previously, each feature corresponds to a PDF of several values with each of them used as a separate predictor. These predictors were combined using a random forest classifier, which is well suited for this category of features [13], as well as the kernel discriminant analysis using spectral regression.

Three classification tasks were defined for this dataset:

- Gender classification. Note that a random classification would predict approximately 50%, as this is a two-class classification.
- Age range classification. To avoid classes with very small patterns, seven age ranges were defined: (1950 to 1965), (1966 to 1975), (1976 to 1985), (1986 to 1990), (1991 to 1995), (1996 to 2000), and (2001 to 2012). A random classification would therefore predict approximately 14%.
- Nationality prediction. To avoid small classes, only writers of eight different nationalities were considered. Each of these classes has more than 30 writers. A random classification would only predict approximately 12%.

Table 3 Correct classification rates for nationality detection using random forests (Rf) and kernel discriminant analysis (Kda)

	Arabic				English				Both			
	Same text (%)		Different text (%)		Same text (%)		Different text (%)		Same text (%)		Different text (%)	
Feature	RF	KDA	RF	KDA	RF	KDA	RF	KDA	RF	KDA	RF	KDA
f1	42.1	39.6	37.5	42.4	35.8	36.4	37.2	39.2	39.4	43.5	38.7	40.3
f2	37.8	38.0	38.2	43.4	36.2	36.4	38.9	41.0	42.6	41.1	40.	37.5
f3	44.0	38.0	39.8	42.8	33.7	36.4	35.2	39.9	38.6	38.6	39.6	37.5
f4	37.1	39.6	37.5	47.7	36.2	34.6	34.7	37.4	37.8	41.1	38.3	44.4
f5	42.1	39.6	39.8	36.3	42.4	37.1	41.8	39.6	43.8	38.2	43.0	41.5
f6	42.5	44.2	42.1	39.7	42.0	44.1	43.9	41.0	45.5	47.2	44.8	43.2
f7	43.2	47.3	42.1	38.6	41.6	45.2	44.8	49.6	44.0	53.7	42.4	48.0
f8	29.3	35.1	28.6	42.8	28.0	33.1	30.1	37.4	30.1	39.0	32.7	35.5
f9	38.2	37.3	33.6	42.2	35.8	33.5	36.4	39.2	34.9	38.6	35.5	37.5
f10	39.4	40.9	38.2	42.2	32.1	36.0	36.4	40.7	36.6	41.1	40.4	36.3
f11	42.1	41.5	41.7	43.4	34.2	40.8	36.0	40.7	38.8	37.8	41.4	39.1
f12	44.8	42.3	44.0	43.4	35.8	41.9	37.2	43.2	40.8	38.6	42.6	37.9
f13	45.6	40.9	44.4	40.9	36.2	45.2	34.3	41.7	42.6	35.4	42.0	38.7
f14	44.4	40.5	43.6	41.4	37.5	40.8	36.4	41.7	40.4	32.1	42.2	39.9
f15	42.1	40.0	42.9	41.1	38.3	42.7	38.5	42.8	43.2	32.5	42.4	40.7
f16	40.2	39.6	43.2	42.2	38.7	40.4	37.2	42.8	42.0	32.9	42.2	38.7
f17	39.4	39.2	43.6	44.5	35.4	40.4	37.7	39.2	42.4	39.4	43.0	39.9
f18	41.3	40.9	36.3	43.7	37.0	34.6	38.1	42.1	39.0	41.1	41.8	38.7
f19	41.3	43.2	36.3	44.1	37.0	35.7	38.1	40.7	39.0	41.9	41.8	37.5
f20	43.2	42.1	37.8	44.7	35.4	40.4	39.8	43.2	42.4	42.3	43.8	40.3
f21	46.0	42.9	38.6	46.4	36.2	40.4	38.9	42.8	42.4	41.5	43.6	38.7
f22	47.1	42.7	40.2	44.3	37.0	40.1	40.2	43.5	43.4	41.5	44.6	39.1
f23	46.3	43.1	44.0	43.7	38.3	43.0	38.5	44.6	44.0	40.7	44.4	39.5
f24	46.0	43.6	44.8	42.4	38.3	44.5	38.9	45.7	44.2	39.0	44.6	37.1
f25	45.6	43.6	45.6	44.9	39.5	44.9	40.2	44.6	44.6	39.8	44.6	37.1
f26	45.2	41.9	43.6	44.3	39.5	43.0	40.6	45.7	44.6	39.8	45.0	37.9
f1 + f2 + f3	41.7	41.3	40.4	45.1	42.3	39.3	45.1	44.2	42.6	43.9	44.0	40.3
f4 +,…, + f7	43.5	47.7	40.8	46.2	47.2	44.5	44.7	45.7	44.5	51.6	41.3	46.0
f8 +,…, + f17	45.7	41.3	46.0	44.3	42.3	44.5	40.7	46.4	44.7	38.2	42.9	41.5
f18 +,…, + f26	46.8	44.4	47.1	45.1	44.0	43.8	39.0	45.3	44.3	41.9	43.1	37.1
f1 +,…, + f26	48.9	47.7	47.4	46.2	44.4	48.9	44.3	47.1	**46.4**	44.3	**46.7**	44.8

4.2 Results

Tables 1, 2, and 3 depicts the correct classification rates for each category of features using a random forest of 5,000 random trees and kernel discriminant analysis for every gender, age range, and nationality classification. The classification is performed for the Arabic and English languages separately in the first step and jointly in the second step. The results are reported for the case of similar texts written by all the writers and different texts for each writer. Figure 7 summarizes the best results for gender, age range, and nationality using two classification methods.

4.3 Discussion and analysis

To test which feature combination is optimal for each classification problem, we plotted the average performance (for similar and different texts using random forest and KDA) for the proposed geometric features (f1 to f3), chain code features (f4 to f7), edge-based directional features (f8 to 17), and filled edge-based directional features (f18 to f26). The results are shown in Figure 8. It is important to note that the performances are seemingly very high for nationality, low for age range and even lower for nationality detection. This is due to the fact that nationality prediction is a binary classification problem in which even a random prediction would score 50%, whereas age range and nationality detection are respectively seven- and eight-class classification problems in which a random classifier would only score 14% and 12%, respectively.

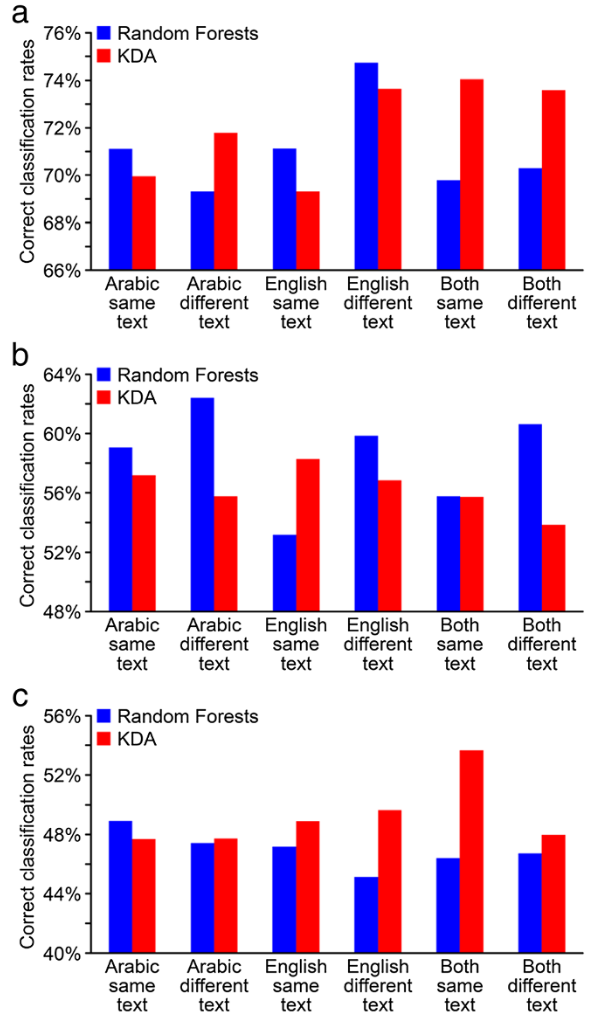

Figure 7 Best correct classification rates using random forests and kernel discriminant analysis. (a) Gender, **(b)** age range, and **(c)** nationality.

The results show that chain code-based features generally outperform using other features for predicting the gender and the nationality which suggests that the detailed distribution of curvatures in the handwriting is of a high importance in characterizing the gender and nationality. Note as well that the proposed geometric features outperforms other features for predicting the age range which suggests that all of the directions, curvatures, and tortuosity are essential for determining the age through handwriting.

We also plotted the average performance of random forests and KDA classifiers when combining all the features (f1 to f26). The results are shown in Figure 9. Random forests are generally preferred for the prediction of age range and nationality, whereas KDA is preferred for the prediction of gender. This clearly suggests that random forests are to be preferred when predicting patterns with many classes whereas KDA are to be preferred for binary classification problems.

The average performance when combining all the features (f1 to f26) on the same and different texts is shown in Figure 10. Notice that handwritings produced by the same writer yield slightly better results for the prediction of gender but not for the prediction of age range or nationality. This suggests that working on the same texts or different texts do not have any benefits in improving the classification results.

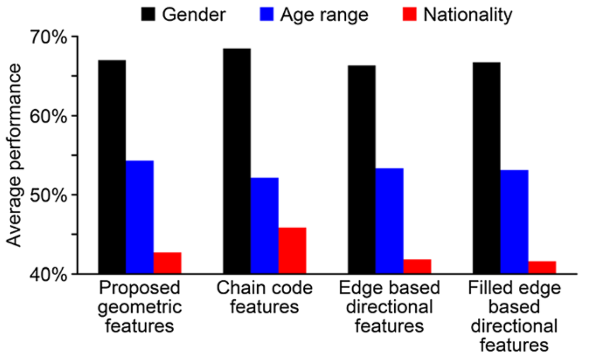

Figure 8 Performance of each feature combination for the detection of gender, age range, and nationality.

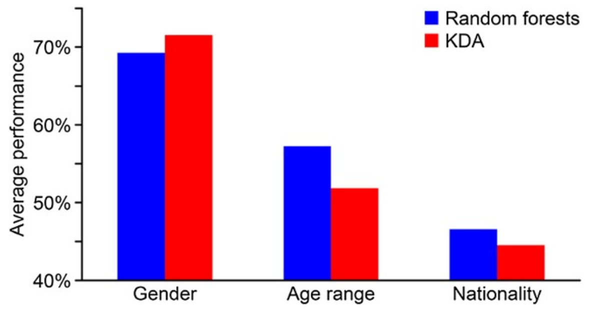

Figure 9 **Average performance of random forests and KDA classifiers for the prediction of gender, age range, and nationality.**

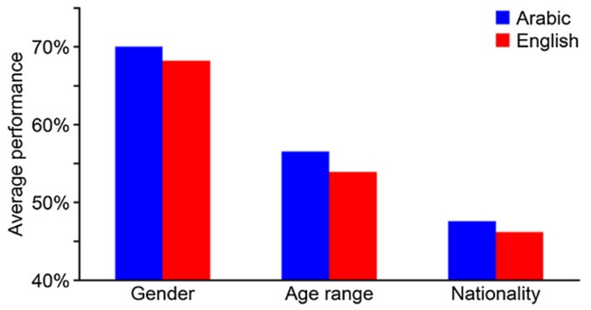

Figure 11 **Average performance for Arabic and English texts.**

The average performance when combining all the features (f1 to f26) on Arabic and English texts is shown in Figure 11. Generally, Arabic handwritings yield better prediction results. This is explained by the complexity of the Arabic script which tends to help better categorize writers.

Additionally, the combination of several features does not always yield better results. There are many cases in which one feature alone outperforms a combination of several features. Indeed, some features might be redundant or irrelevant and contain no useful information in which case they need to be removed for obtaining better performance.

The classification systems described here are promising; however, there remains a lot of room for improvement in terms of using new features and classification methods. Comparison of results, obtained in this research, with other researchers is difficult because of differences in experimental details, the actual handwriting used, the method of data collection, and dealing with cursive off-line handwritten text. If this work is compared to writer demographic identification research [1,4], it is the first one that implemented on offline cursive Arabic and English writers. This also means that it uses different sets of features and classification techniques. Unfortunately, both datasets used in [1,4] are not publically available. The dataset used in this research is available for research purposes.

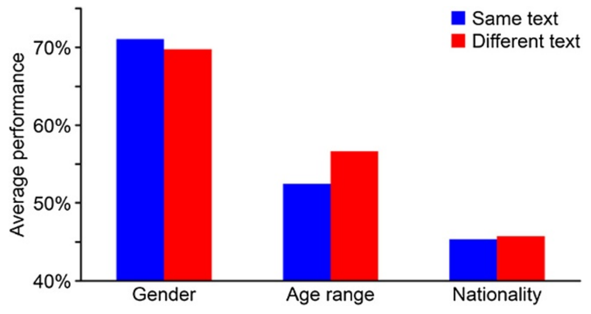

Figure 10 **Average performance for same and different texts.**

Finally, for the comparison purposes, the average correct gender classification results are over 73%, which exceeds the results reported in [4] for offline gender identification (55.39%) on a different dataset consisting of 200 writers. The results also compare well with the 77.5% reported in [1] on a smaller dataset (800 individuals wrote the same letter). The authors of [1] also report an age range classification accuracy of 86.6%, which seemingly outperforms our 55%. However, the authors only included two age range categories (below 24 and above 45) and included only 650 individuals.

5 Conclusion

We have presented a method that uses several geometric features for the classification of age range, gender, and nationality of handwritings, which is applicable for both Arabic and English documents. This study is the first that reported classification results for those subcategories on the QUWI dataset [19]. The results are reported for both text-dependent and text-independent category classification.

Experiments show that using chain code-based features generally outperforms using other features for predicting the gender and the nationality, and the proposed geometric features outperforms other features for predicting the age range. The results suggest that random forests are generally preferred for the prediction of age range and nationality, whereas KDA is preferred for the prediction of gender. We have also noticed that handwritings produced by the same writer yield slightly better results for the prediction of gender but not for the prediction of age range or nationality. It has also shown that experiments on Arabic handwritings attained generally better prediction results. Future work includes exploring ways of combining the proposed features and using other classifiers. The use of the proposed features for predicting the handedness of writers is also planned.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

This work is supported by the Qatar National Research Fund through National Priority Research Program (NPRP) No. 09 – 864 – 1 – 128. The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official views of the Qatar National Research Fund or Qatar University.

Received: 13 October 2013 Accepted: 29 January 2014
Published: 17 February 2014

References

1. K Bandi, SN Srihari, Writer demographic identification using bagging and boosting, in *Proceedings of the International Graphonomics Society Conference (IGS)* (Salerno, Italy, 2005), pp. 133–137. 26–29 June
2. S Srihari, SH Cha, H Arora, S Lee, Individuality of handwriting: a validation study, in *2001 Proceedings of the Sixth International Conference on Document Analysis and Recognition* (Seattle, 2001), pp. 106–109. 10–13 September
3. M Liwicki, A Schlapbach, P Loretan, H Bunke, Automatic detection of gender and handedness from on-line handwriting, in *Proceedings of the 13th Conference of the International Graphonomics Society* (Melbourne, 2007), pp. 179–183. 11–14 Novembers
4. M Liwicki, A Schlapbach, H Bunke, Automatic gender detection using on-line and off-line information. Pattern. Anal. Appl. **14**, 87–92 (2011)
5. N Otsu, A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man Cybern. **9**(1), 62–66 (1979)
6. A Hassaine, S Al-Maadeed, J Alja'am, A Jaoua, A Bouridane, The ICDAR2011 Arabic Writer Identification Contest, in *Proceedings of the Eleventh International Conference on Document Analysis and Recognition* (Beijing, China, 2011). 18–21 September
7. A Hassaïne, S Al-Maadeed, A Bouridane, A set of geometrical features for writer identification, in *The 19th International Conference of Neural Information Processing Doha, Qatar* (Springer, Berlin Heidelberg, 2012), pp. 584–591. 12–15 November
8. K Koppenhaver, *Forensic Document Examination: principles and practice* (Humana Press, New York, 2007)
9. M Bulacu, L Schomaker, Text-independent writer identification and verification using textural and allographic features. IEEE Trans. Pattern Anal. Mach. Intell. **29**(4), 701–717 (2007)
10. J Matas, Z Shao, J Kittler, Estimation of curvature and tangent direction by median filtered differencing, in *The 8th International Conference on Image Analysis and Processing*. Lecture notes in computer science. vol 974 (Springer-Verlag, Berlin, 1995), pp. 83–88. 13–15 September
11. TY Zhang, A fast parallel algorithm for thinning digital patterns. Commun. ACM **27**(3), 236–239 (1984)
12. I Siddiqi, N Vincent, Text independent writer recognition using redundant writing patterns with contour-based orientation and curvature features. Pattern Recogn. **43**(11), 3853–3865 (2010)
13. L Breiman, Random forests. Mach. Learn. **45**, 5–32 (2001)
14. D Cai, X He, J Han, *Proceedings of the ICDM* (Omaha, Nebraska, 2007). 28–31 October
15. HH Bock, E Diday, *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data* (Springer, Heidelberg, 2000)
16. A Liaw, M Wiener, Classification and regression by randomforest. NANR News **2**(3), 18–22 (2002). http://CRAN.R-project.org/doc/Rnews
17. S Mika, G Ratsch, J Weston, B Scholkopf, KR Mullers, Fisher discriminant analysis with kernels, in *Neural Networks for Signal Processing IX, 1999.* Proceedings of the 1999 IEEE Signal Processing Society Workshop, Madison (IEEE, Piscataway, 1999), pp. 41–48. 23–25 August
18. TY Lin, Y Xie, A Wasilewska, CJ Liau, *Data Mining: Foundations and Practice, vol. 118* (Springer, Heidelberg, 2008)
19. S Al-Ma'adeed, W Ayouby, A Hassaine, J Aljaam, QUWI: an Arabic and English handwriting dataset for offline writer identification, in *International Conference on Frontiers in Handwriting Recognition* (Bari, Italy, 2012). 18–20 September

doi:10.1186/1687-5281-2014-10
Cite this article as: Al Maadeed and Hassaine: **Automatic prediction of age, gender, and nationality in offline handwriting.** *EURASIP Journal on Image and Video Processing* 2014 **2014**:10.