Pattern Anal Applic (2016) 19:699–708
DOI 10.1007/s10044-014-0438-y

THEORETICAL ADVANCES

Novel geometric features for off-line writer identification

Somaya Al-Maadeed · Abdelaali Hassaine · Ahmed Bouridane · Muhammad Atif Tahir

Received: 15 September 2013 / Accepted: 16 November 2014 / Published online: 27 December 2014

Abstract Writer identification is an important field in forensic document examination. Typically, a writer identification system consists of two main steps: feature extraction and matching and the performance depends significantly on the feature extraction step. In this paper, we propose a set of novel geometrical features that are able to characterize different writers. These features include direction, curvature, and tortuosity. We also propose an improvement of the edge-based directional and chain code-based features. The proposed methods are applicable to Arabic and English handwriting. We have also studied several methods for computing the distance between feature vectors when comparing two writers. Evaluation of the methods is performed using both the IAM handwriting database and the QUWI database for each individual feature reaching Top1 identification rates of 82 and 87 % in those two datasets, respectively. The accuracies achieved by Kernel Discriminant Analysis (KDA) are significantly higher than those observed before feature-level writer identification was implemented. The results demonstrate the effectiveness of the improved versions of both chain-code features and edge-based directional features.

S. Al-Maadeed (✉) · A. Hassaine
Computer Science and Engineering Department,
College of Engineering, Qatar University, Doha, Qatar
e-mail: s_alali@qu.edu.qa

A. Hassaine
e-mail: hassaine@qu.edu.qa

A. Bouridane · M. A. Tahir
Department of Computer Science and Digital Technologies,
Northumbria University, Newcastle upon Tyne, UK
e-mail: ahmed.bouridane@northumbria.ac.uk

M. A. Tahir
e-mail: muhammad.tahir@northumbria.ac.uk

Keywords Forensic document examination · Writer identification · Handwriting curvature · Handwriting direction · Handwriting tortuosity

1 Introduction

Automatic writer identification is important in forensic document examination. Numerous cases have dealt with evidence provided by handwritten documents, such as wills and ransom notes [1]. Moreover, writer identification can be used in handwriting recognition when adapting the recognizers to a specific type of writers [2] and in handwriting synthesis when generating a text as it would have been written by a specific writer [3]. Writer identification methods generally consist of two main steps. The first step is feature extraction, in which discriminating features are extracted from the handwritten documents to be compared while the second step involves matching or classification in which a comparison between the features is computed, and a decision regarding the authorship is made according to the distance between the extracted features. What makes a system powerful and robust is closely related to a strong feature extraction step because the extraction of discriminative features helps to distinguish between writers.

Automatic methods for writer identification can be classified into two main categories: codebook-based and feature-based approaches. In codebook-based approaches, the writer is assumed to act as a stochastic generator of graphemes. The probability distribution of a grapheme is a characteristic of each writer and can efficiently be used to distinguish between different writers. The methods in this category mainly differ in how the handwriting is segmented into graphemes and how the graphemes are clustered. On the other hand, feature-based approaches

compare the handwriting samples according to geometrical [4], structural [5], or textural features [6, 7]. Feature-based approaches are proven to be efficient and are generally preferred when only a limited amount of handwriting data is available.

In the remainder of this section, an overview of the works done in the field of off-line writer identification is presented.

Srihari et al. proposed a set of macro-features that are extracted from a document, paragraph, or word level (i.e., entropy of gray values; gray-level threshold; number of black pixels; number of interior and exterior contours; number of vertical, horizontal, negative, and positive slope components; and slant) and micro-features that are extracted at the word or character level (i.e., gradient, structural, and concavity features). This approach has been validated on two datasets of 711 writers using the same letter [1] achieving a classification performance of 89 and 87 %, respectively.

Said et al. described a text-independent writer identification method based on Gabor filtering and grayscale co-occurrence matrices. The authors obtained 95 and 88 % recognition rate, but evaluated on a small set of 10 and 15 writers [7].

Marti et al. used text line-based features for text-independent writer identification [5]. The authors used features related to the position of the top line, the bottom line, the upper baseline, and the lower baseline; the width; and the slant of the writing. They also introduced a set of features based on fractal geometry to distinguish between badly formed and legible writing. The recognition rate obtained with their method reached 90 % on a subset of 50 writers from the IAM database [8]. The IAM database will also be used in the present study though more writers will be considered.

A further study by Hertel and Bunke [9] resulted in novel features based on connected components, enclosed regions, lower and upper contours, fractal features, and basic features (including writing skew and slant, the height of the three main writing zones, and the width of the writing). This study has been validated on a subset of 50 writers of the IAM database. A recognition rate of 99 % was obtained for this small data set. Connected component feature alone obtained 31 % in the same dataset.

Schlapbach and Bunke used the results of the HMM text recognizer for both the identification and verification of writers [10]. The authors proposed recognizers that are trained for each individual writer. Subsequently, a handwritten sample in question is provided as input to all the recognizers, which the corresponding output likelihood ratio values. By sorting these values, the identity of the most probable writer then is determined. This study has validated their approach using 100 writers in the IAM database with a recognition rate of 96 %. In a follow-up study, the authors proposed an improvement by 2 % to the system by deploying Gaussian mixture models [11].

Bensefia et al. used a textual-based information retrieval model for the writer identification stage [12]. This makes it possible to use a particular feature space based on feature frequencies of occurrence. Image queries are handwritten documents projected in this feature space. This approach achieved an 86 % identification rate on a subset of 150 documents of the IAM database.

Siddiqi and Vincent proposed the use of the redundancy of graphemes to characterize writer individuality [13]. The authors also proposed some chain code-based features extracted from the handwriting contours. Their best performing feature (local stroke direction histogram) achieved a 77 % identification rate with the IAM database [14]. The two categories of features were combined in a follow-up study [15] and validated with the IAM database. Their best performing feature is a distribution of chain code with achievement of 79 %. The chain-code feature initially introduced by these authors will be further improved in this study.

Bulacu and Schomaker used edge-based directional probability distribution functions (PDFs) as features for text-independent writer identification. The joint PDF of "hinged" edge–angle combinations outperformed all the other evaluated features (including Contour-direction PDFs, Direction co-occurrence PDFs, Grapheme emission PDFs, Run length on background PDFs, and Autocorrelation during horizontal raster scanning) [16]. By using Contour-hinge PDF features a recognition rate of 81 % was achieved.

A study by Imdad et al. used steered Hermite features combined with the SVM classifier [17]. The proposed method achieved an 83 % identification rate on a subset of 30 writers from the IAM database.

Muda et al. showed that the discretization of features can significantly improve the identification rates [18]. Discretization is performed by exploring the partitioning of features into intervals and unifying the values for each interval. The method achieved around 99 % but 60 writers only were used in the evaluation.

Bozeková combined grapheme features obtained using a Kohonen Self-Organizing Map and specified structural features to achieve a 96.5 % identification rate on a subset of 40 writers from the IAM database [19].

Dolega et al. used derived canonical stroke frequency descriptors from handwritten text to identify writers [20]. The authors reached an 88 % identification rate on a subset of 50 writers in the IAM database.

Steinke et al. combined local features that use different mathematical procedures, such as the reproduction of the write line of individual characters by Legendre

polynomials, and global textural features [21]. The proposed method achieved a 99.5 % identification rate on a subset of 93 writers. Jain and Doermann extracted codebooks of K-adjacent segments from handwriting text to characterize writer individuality [22]. The method achieved a 93 % identification rate on 300 writers from the IAM dataset.

From the review and analysis of existing works available in the literature it can be concluded that existing methods are usually validated using a small number of writers which are usually less than 100, whereas the IAM dataset contains more than 650 writers. In addition, one can note that there has been a consistent amount of research on the combination of features; however the performances of individual features have not been studied and need further analysis with a view to determine their discriminative power to maximize their recognition performances and also to aid as to how one can combine them efficiently. Another motivation of the work relates to the development of novel features especially to capture the peculiar characteristics of the writers' handwriting relating to direction, curvature and also tortuosity. To address the above issues, this paper has the following contributions:

- New features based on direction (f1), curvature (f2), tortuosity (f3) quadruple-order chain code (f7) are proposed including an implementation and evaluation on much recent large databases. To the best of the authors' knowledge, the proposed technique is the first attempt the field of handwriting identification.
- The paper also proposes an improvement of state-of-the-art edge-based directional features (by using a filled moving window instead of an edge moving window) and the chain code-based features (by using a fourth-order chain-code list) in terms of the discriminative power (f18–f26). We demonstrate that these improvements lead to much enhanced identification rates.
- A kernel discriminant analysis has been used in order to combine the features using several metrics including X^2, L^2, and L^1 distances to evaluate the performances. An analytical study is then given showing improvement in the performances.

The remainder of the paper is organized as follows: Sect. 2 introduces the proposed methodology and the system's components including the description of the features used and the matching strategy used. In this section a set of novel features are also discussed. Section 3 discusses the datasets used in our experimentation and evaluation. Section 4 presents the results and their analysis while Sect. 5 concludes this paper.

2 Proposed methodology

Similar to any writer identification system, our proposed method consists of two steps: feature extraction and matching steps. In the following we will discuss both steps in detail.

2.1 Feature extraction

In this stage, the characterizing features are extracted from the handwriting. To develop a pen-independent system, images are first binarized using the Otsu thresholding algorithm [25]. It is worth noting that in writer identification, features do not correspond to a single value but to a probability distribution function (PDF) extracted from the handwriting images to characterize a writer's uniqueness [26]. Novel sets of features were extracted from different handwritten datasets. In this paper we propose to extract features based on direction (which we refer to as f1), curvature (f2), tortuosity (f3), quadruple-order chain code (f7), and the edge-based directional features using the whole window computed from size 2 (f18, whose PDF size is 12) to size 10 (f26, whose PDF size is 220). We have also considered and used other state-of-the-art features, such as chain-code features (f4–f6) and Edge-based directional features (f8–f17) to compare the results. These features will be explained in the what follows.

2.1.1 Direction (f1)

Direction is known to be useful for characterizing writers [16]. The methodology used in this paper is novel and has been used before in writer identification. Its implementation is somehow similar to the method proposed by Matas et al. [27]. First, we compute the Zhang skeleton of the binarized image. This skeleton is well known for not producing parasitic branches in contrast to most skeletonization algorithms [28]. The skeleton is then segmented at its junction pixels. Then, we move along the pixels of the obtained segments of the skeleton using the predefined order, favoring the four-connectivity neighbors, as shown in Fig. 1a. A result of such an ordering is shown in Fig. 1b. For each pixel p, we consider the $2N + 1$ neighboring pixels centered at position p. The linear regression of these pixels gives a good estimation of the tangent at pixel p (Fig. 1c). The value of N has empirically been set to 5 pixels. This feature is illustrated in Fig. 1d.

The PDF of the resulting directions is computed as a vector of probabilities for which the size has been empirically set to 10. Note that this way of computing directions has never been used before in off-line writer identification.

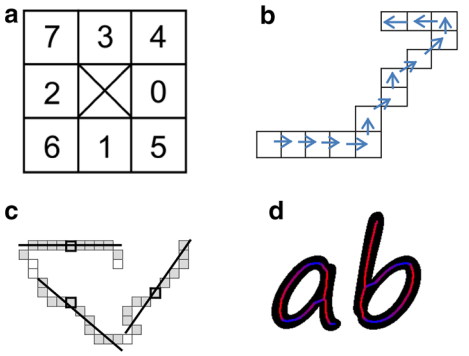

Fig. 1 Computing local direction. **a** The predefined order for traversing shapes. **b** Example of an ordered shape. **c** Estimating direction by computing the linear regression of neighboring pixels for three different *bold-colored regions*. **d** Binary image and its corresponding Zhang skeleton; the *red color* corresponds to a $\pi/2$ tangent, and the *blue color* corresponds to a zero tangent (color figure online)

2.1.2 Curvature (f2)

Curvature is commonly accepted in forensic document examination as an important discriminative feature [29]. Here, we have introduced a novel method for computing curvature and its deployment in the field of writer identification for the first time. It is to ne noted that this technique has been used previously for estimating the curvature of the peaks and valleys in optical soundtracks [30]. We have modified and adapted this method to handwritings 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 n_2 which belong to the background and the foreground, respectively (Fig. 2a). Therefore, the difference $n_1 - n_2$ increases with the local curvature of the contour. Therefore, we estimate the curvature as follows:

$$C = \frac{n_1 - n_2}{n_1 + n_2}.$$

This value is illustrated in Fig. 2b on a binary shape for which t has been empirically set to 5.

The PDF of curvature is computed in a vector whose size has been empirically set to 100. To the best of our knowledge, this method of computing curvature is also novel in the field of off-line writer identification.

2.1.3 Tortuosity (f3)

In this work we propose to use tortuosity as a novel feature to allow us 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 (Fig. 3a). An example of estimated tortuosity is shown in Fig. 3b.

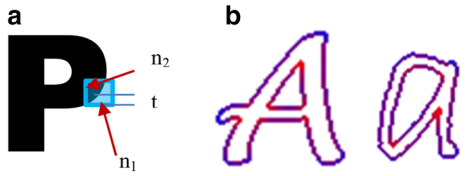

Fig. 2 **a** Computing curvature. **b** Curvature highlighted in the *binary image*; *red* corresponds to the maximum curvature, and *blue* corresponds to the minimum curvature (color figure online)

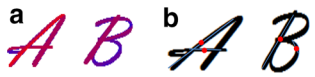

Fig. 3 Computing tortuosity: **a** longest traversing segment for four different pixels. **b** Length of maximum traversing segment; *red* corresponds to the maximum length and *blue* to the minimum length (color figure online)

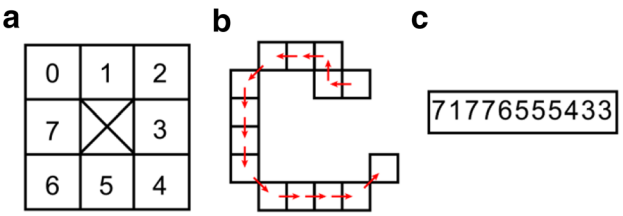

Fig. 4 **a** Order followed to generate the chain code. **b** Example shape and **c** its corresponding chain code

The PDFs of the angles of the longest traversing segments are stored in a vector whose size has been set to 10.

2.1.4 Chain-code features (f4–f7)

Chain codes are generated by browsing the contour of the text and assigning a number to each pixel according to its location with respect to the previous pixel. Figure 4 shows a contour and its corresponding chain code.

These features make it possible to characterize the detailed distribution of curvature in the handwritings. Chain codes can be applied at different levels:

f4: 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: 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.

Similarly, **f6** and **f7** correspond to PDFs of (i, j, k) and (i, j, k, l) in the chain-code list, and their respective sizes are 512 and 4,096, respectively.

Although, the chain-code features f4, f5, and f6 have previously been applied to writer identification [14], we propose to deploy the quadruple-order chain code f7 for the first time in writer identification.

2.1.5 Edge-based directional features (f8–f26)

Initially introduced in [16], these features provide detailed distributions for the direction and can also be applied at several scales by positioning a window centered at each contour pixel and counting the occurrences of each direction, as shown in Fig. 5a. This feature has been computed from size 1 (**f8**, whose PDF size is 4) to size 10 (**f17**, whose PDF size is 40). We have also extended these features to include not only the contour of the moving window but also of the whole window (Fig. 5b). This feature has been computed from size 2 (**f18**, whose PDF size is 12) to size 10 (**f26**, whose PDF size is 220).

2.2 Matching of feature vectors

When comparing a query document q against any given document i, the difference between their features is computed as shown in Fig. 6.

In this study, three different distances were considered and evaluated:

- The χ^2 distance is the first method used for comparing two PDFs: $\chi^2 = \sum_{n=1}^{\text{size}} \frac{\left(F_q(n)-F_i(n)\right)}{F_q(n)+F(n)_i}$
- The L^2 distance or Euclidian distance was also tested: $L^2 = \sum_{n=1}^{\text{size}} \left(F_q(n) - F_i(n)\right)^2$
- The L^1 distance was also tested: $L^1 = \sum_{n=1}^{\text{size}} \left|F_q(n) - F_i(n)\right|$

In addition to computing one distance between each pair of feature vectors, we have also computed the list of distances at the element level for each pair of vectors:

$$\text{list of distances} = \{F_q(n) - F_i(n) \text{such as}: n \in \{1, 2, \ldots, \text{size}\}\}$$

These distances are combined using a kernel discriminant analysis with spectral regression (SR-KDA). A description of this classifier is given below:

Let $x_i \in Rd$, $i = 1,\ldots$ and m be training vectors represented as an $m \times m$ kernel

Matrix K is defined such that $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$, where $\Phi(x_i)$ and $\Phi(x_j)$ are the embedded data items x_i and x_j, respectively. If ν denotes a projective function into the kernel feature space, then the objective function for KDA is

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

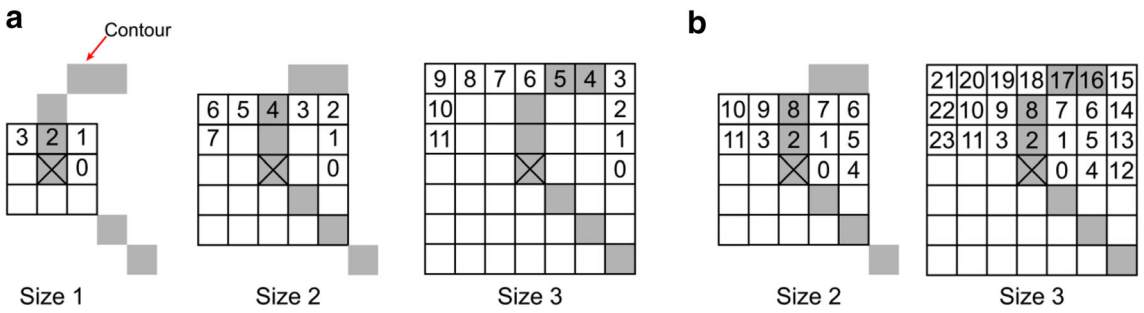

Fig. 5 Counting the edge-based directional features when considering **a** the contour of the moving window and **b** the whole moving window

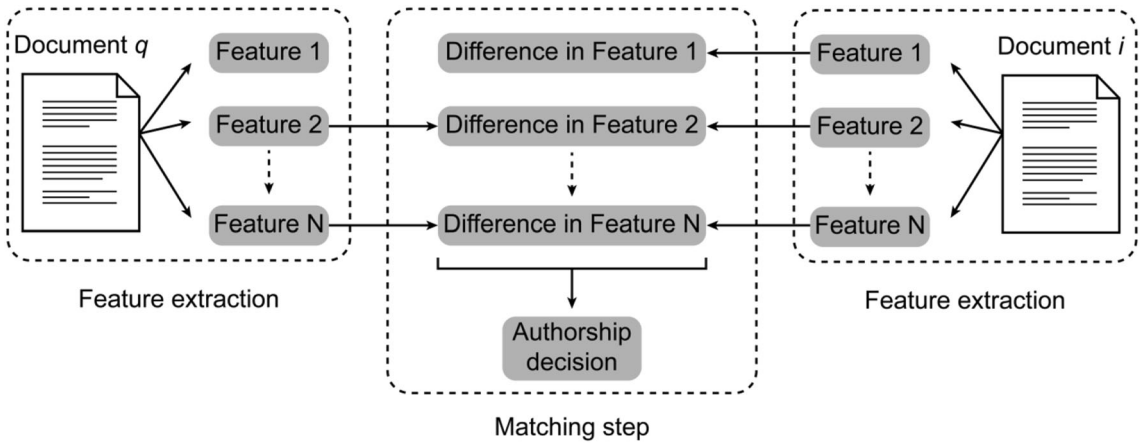

Fig. 6 General scheme of the proposed method

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 eigenvalue problem $C_b = \lambda C_t$. It is proven that Eq. (1) is equivalent to the following:

$$\max_{\alpha} D(\alpha) = \frac{\alpha^T K W K_\alpha}{\alpha^T K K_\alpha} \tag{2}$$

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

$W = (W_1)_{l=1,\ldots,n}$ is a $(m \times m)$ block diagonal matrix of labels arranged such that the upper block corresponds to positive examples and the lower one to negative examples of the class. Each eigenvector α gives a projection function ν into the feature space.

Instead of solving the eigenvalue problem in KDA, the KDA projections can be obtained by the following two linear equations:

$$\begin{aligned} W\phi &= \lambda\phi \\ (K + \delta I)\alpha &= \phi \end{aligned} \tag{3}$$

where is an eigenvector of W, I is the identity matrix, and $\delta > 0$ is a regularization parameter. $W = (W_1)_{l=1,\ldots,n}$ is an $(m \times m)$ block diagonal matrix of labels arranged such that the upper block corresponds to positive examples and the lower one to negative examples of the class. Eigenvectors ϕ are obtained directly from the Gram-Schmidt method. Because $(K + \delta I)$ is positive definite, the Cholesky decomposition is used to solve the linear equations in (3). Thus, for the resolution of linear system (3), the system becomes

$$K * \alpha = \phi \Leftrightarrow \begin{cases} R^T\theta = \phi \\ R_\alpha = \theta \end{cases} \tag{4}$$

i.e., the system is solved first to find vector θ and then vector α. In summary, SRKDA only needs to solve a set of regularized regression problems, and there is no eigenvector computation involved. This approach results in substantial improvement in terms of computational cost and makes it possible to handle large kernel matrices. After obtaining α, the decision function for new data items is calculated from $f(x) = \sum_{i=1}^{n} \alpha_i K(x, x_i)$, where $K(x, x\_i) = \langle \Phi(x), \Phi(x\_i) \rangle$.

The proposed features have different discriminative levels, which are described below.

3 Training and testing datasets

For the experimental set up of our system, we have chosen two different datasets: IAM [8] and QUWI [23]. The IAM dataset contains English handwritings whereas the QUWI dataset contains both English and Arabic handwritings. The IAM dataset [8] is the most widely used dataset in writer identification. It contains handwritings of 657 writers, each of whom produced two samples of text. The IAM database consists of forms with handwritten English text of variable content that has been scanned at 300 dpi with 8 bits/pixel in grayscale. In addition to writer identity, the images are accompanied by extensive segmentation and ground-truth information at the text line, sentence, and word levels. This dataset includes a variable number of handwritten pages per writer, from one page (350 writers) to 59 pages (one writer). To have comparable experimental conditions with existing state-of-the-art methods, we have modified the IAM dataset to contain two samples per writer. We kept only the first two documents for the writers who produced more than two documents, and we split the document roughly in half for the writers with a unique page in the original set. This modified IAM is used for testing purposes in a manner similar to that used previously [14–16]. This modified dataset contains lowercase handwritings from 657 people, with two samples per writer. The amount of ink is roughly equal in the two samples belonging to one writer but varies among writers from three lines up to a full page. Note that the third and fourth documents of the writers who produced at least four documents are used for training purposes.

We also used the QUWI dataset [23], which is a dataset built at Qatar University consisting of handwritings from 1,017 writers. This dataset has been scanned with a spatial resolution of 600 dpi. The writers were asked to produce four pages of text: one similar text and one different text in English and one similar text and one different text in Arabic. Part of the QUWI dataset has been used for organizing a competition for writer identification [24].

4 Evaluation

The most widely used database for writer identification is the IAM database [8] which was described above. Similar to the method described in [16], we have used only the first two documents from the writers who produced more than two documents, and for the writers who produced only one document, we split the document into two separate documents. Note that the comparison with other systems is to be considered approximate and not exact because the current IAM database contains 657 writers, not 650.

Therefore, each document is compared against all 1,313 other documents with only one possible correct match. If the distance between the document of interest and the correct match is the smallest among all possibilities, then the document is said to be correctly identified. The TOP 10 identification rate considers the matching as correct if the corresponding distance is among the 10 smallest distances.

As previously stated in Sect. 3, the other dataset used in this study is the QUWI dataset. The pages of this dataset have been segmented into three paragraphs; two of them are used for training while the third one is used for testing. The identification rates obtained for the presented features for both datasets where tested using the χ^2 distance, L^2 distance, L^1 distance, and KDA classifier. The results obtained for each feature using the three types of distances are given in Table 1.

The results of the KDA classifier on the test datasets are presented in Table 2.

As mentioned previously, in this paper, the proposed features were computed as follows: directional feature is determined using a linear regression technique (f1) while the curvature (f2), tortuosity (f3), quadruple-order chain code (f7) and edge-based directional features are computed using the whole window from a size 2 (f18, whose PDF size is 12) to size 10 (f26, whose PDF size is 220). Other chain-code features (f4 to f6) and direction features (f8 to f17) were reproduced for comparison purposes.

The KDA classifier achieves the best TOP1 recognition rate for most features except the following: f2, f16, and f17 in all datasets; f8 in all QUWI datasets; and f7 and f9 in the QUWI Arabic same-text dataset. The TOP10 recognition rates using kDa classifier are also generally the highest in the IAM dataset.

In the case of the QUWI dataset, the highest recognition rates were obtained using the L^1 distance. It was also found that the first 17 features give high performances for most sub-datasets when using the other distances. The KDA classifier produced the highest recognition rate for the 10 highest-ranked writers for the QUWI dataset for features f18 to f26.

For the IAM dataset, the quadruple-order chain-code feature f7 yielded the highest recognition rate of 82 % for the top writer and of 92 % for the top ten writers using the

Table 1 Top1 identification rates using each category of distance

Feature	IAM			QUWI Arabic different text			QUWI Arabic same text			QUWI English different text			QUWI English same text		
	χ^2	L^2	L^1	χ^2	L^2	L^1	χ^2	L^2	L^1	χ^2	L^2	L^1	χ^2	L^2	L^1
f1	40.56	35.31	36.30	13.98	12.60	12.99	30.18	26.63	26.73	18.40	15.80	18.09	23.35	20.17	21.34
f2	36.15	27.93	32.19	20.37	15.85	17.42	30.47	22.49	25.74	27.65	20.06	21.62	34.82	27.71	31.21
f3	42.24	35.24	36.30	15.45	11.81	13.19	30.08	24.65	26.43	18.92	14.35	15.38	25.48	19.85	20.38
f4	32.19	0.15	30.14	15.26	14.07	14.86	27.12	26.43	25.84	21.31	18.92	19.54	28.66	26.65	27.39
f5	59.74	44.60	52.51	43.60	27.36	35.43	69.03	47.53	59.66	51.98	34.20	43.45	62.10	41.61	52.55
f6	**70.47**	51.67	64.92	53.84	31.99	46.26	76.92	54.64	71.30	60.08	37.53	53.74	72.40	46.28	65.18
f7	71.61	55.33	69.48	46.95	33.07	49.90	69.82	54.24	73.27	53.64	38.25	**56.13**	**72.93**	47.03	**68.79**
f8	16.74	15.83	15.75	6.10	6.00	6.30	10.95	10.75	10.65	9.77	9.46	9.46	12.00	11.36	10.83
f9	41.55	36.83	39.04	22.74	21.26	21.36	40.63	39.55	38.26	27.03	25.78	25.57	30.68	28.56	28.66
f10	49.32	0.15	47.56	31.79	30.51	29.33	48.32	46.75	45.27	30.77	29.42	29.00	39.60	38.32	40.13
f11	52.21	50.46	51.07	30.22	29.23	29.33	48.32	46.84	45.96	31.29	30.04	31.29	37.26	36.41	36.94
f12	53.50	52.66	52.05	27.66	27.26	25.89	45.27	43.10	43.39	28.27	26.82	27.13	33.65	31.63	31.85
f13	51.29	48.78	50.84	27.07	26.18	25.30	45.27	43.59	43.29	25.16	24.43	24.43	30.47	29.09	30.25
f14	50.46	49.09	49.39	26.67	25.59	24.51	41.72	39.45	40.14	24.95	24.43	24.53	28.34	27.92	27.92
f15	48.71	46.80	47.72	24.31	21.46	22.64	39.35	37.57	37.67	24.32	24.12	23.70	28.24	27.60	27.71
f16	49.09	47.11	48.86	22.64	21.46	20.77	39.74	36.88	36.88	22.35	21.62	22.87	27.18	26.01	25.69
f17	47.56	45.43	47.03	21.36	21.36	20.96	37.87	35.11	34.42	22.14	20.79	21.73	26.33	25.69	25.27
f18	42.92	37.37	38.13	30.51	27.66	27.36	52.96	50.89	49.51	35.45	32.22	31.60	40.45	38.64	37.16
f19	54.79	48.78	50.53	44.29	42.52	40.94	66.96	66.86	64.50	48.44	46.78	44.49	56.37	54.46	54.25
f20	59.74	55.86	57.84	49.31	47.74	47.24	70.61	70.61	70.12	54.37	51.98	51.46	61.36	60.93	61.47
f21	63.47	60.20	62.18	**51.57**	50.69	**49.70**	71.40	71.60	70.71	54.37	53.74	51.87	64.01	61.89	62.42
f22	64.76	62.02	63.70	50.00	51.08	49.21	71.70	72.49	70.61	55.82	55.30	53.95	64.23	62.85	63.59
f23	67.28	64.16	66.51	49.31	50.79	48.72	72.98	73.18	70.51	55.51	55.41	54.78	64.44	63.80	63.06
f24	68.04	65.83	67.28	48.52	**51.18**	48.03	73.57	74.16	72.49	55.20	56.76	53.64	65.18	64.23	62.95
f25	68.87	66.90	68.80	48.23	50.98	47.74	**72.98**	**75.44**	71.79	56.03	56.03	53.95	64.86	65.29	63.69
f26	70.40	**67.43**	**70.02**	47.64	50.69	47.24	72.49	74.85	**71.99**	**56.96**	**57.80**	54.16	65.50	**65.82**	63.91

Bold values indicate the best results obtained

Table 2 Identification rates of the presented features using the KDA classifier

Feature	IAM		QUWI Arabic diff text		QUWI Arabic same text		QUWI English diff text		QUWI English same text	
	TOP1 (%)	TOP10 (%)	TOP1 (%)	TOP10 (%)	TOP1 (%)	TOP10 (%)	TOP1 (%)	TOP10 (%)	TOP1 (%)	TOP10 (%)
f1	46.70	76.70	17.52	37.30	32.64	56.90	21.21	44.39	27.07	53.40
f2	21.60	36.70	21.75	38.68	25.74	46.25	20.58	43.66	29.09	49.89
f3	47.50	79.10	19.69	39.57	34.91	61.83	23.49	46.78	32.70	57.54
f4	39.30	71.20	19.78	42.13	29.68	59.27	25.99	51.87	32.38	58.92
f5	71.20	85.80	52.46	75.89	72.39	89.84	61.64	81.29	74.73	90.98
f6	78.50	88.70	58.46	77.07	77.42	90.34	63.41	80.67	78.87	91.51
f7	**82.70**	**92.20**	55.81	74.11	71.50	86.49	62.99	80.46	79.09	90.45
f8	18.20	58.20	6.40	20.96	9.76	32.84	8.52	27.86	11.25	31.74
f9	50.30	79.30	25.79	50.30	37.28	67.16	27.44	55.20	31.85	60.83
f10	57.00	81.90	36.32	59.84	49.70	78.99	37.53	66.01	44.27	69.43
f11	57.30	79.10	36.81	62.30	53.85	76.63	38.15	63.83	43.21	70.91
f12	56.50	77.30	35.73	59.65	50.39	74.46	33.37	59.04	41.08	64.86
f13	53.30	78.30	35.14	55.91	49.01	72.78	29.63	54.16	35.99	62.42
f14	53.70	76.60	32.78	54.23	47.14	70.41	30.56	52.81	34.08	59.66
f15	50.30	74.10	30.02	50.59	44.87	69.23	29.11	50.73	33.44	59.45
f16	47.90	71.20	27.85	46.85	43.59	66.17	26.82	49.06	29.30	55.94
f17	47.00	69.00	28.35	47.93	42.50	65.09	24.12	43.76	29.62	52.87
f18	58.10	81.80	40.16	66.34	57.00	82.74	44.80	70.58	50.11	78.87
f19	69.20	87.10	57.68	80.31	76.04	92.90	61.33	84.51	72.61	88.22
f20	72.80	87.10	65.75	84.84	82.74	95.56	71.21	88.15	77.39	91.61
f21	74.30	89.20	68.01	86.02	84.22	95.56	74.01	89.09	80.68	92.36
f22	76.30	89.00	69.29	86.52	85.21	95.86	75.26	**90.75**	81.95	93.21
f23	77.40	89.80	69.88	86.61	86.39	95.56	75.26	90.23	83.01	93.42
f24	78.20	89.70	**70.08**	**86.22**	**87.18**	**95.56**	**76.72**	90.02	82.59	93.52
f25	78.80	89.70	70.08	86.02	87.67	96.06	76.61	89.81	82.38	**93.63**
f26	79.10	90.20	70.08	85.63	**87.67**	**96.25**	77.03	89.40	**82.70**	93.52

Bold values indicate the best results obtained

KDA classifier. This approach marginally outperformed the approaches described in [15, 16] for other individual features using the same dataset. The proposed new edge-based directional features (f16 to f18) resulted in high recognition rates as well. Other individual features produced reasonable recognition rates for the IAM dataset.

In the case of the QUWI dataset, the highest recognition rate was obtained when using the new edge directional feature for English and Arabic for both text-dependent and text-independent approaches. Recognition rates of 70 % for text-dependent Arabic handwriting for the top writer and 86 % for the top 10 ranked writers were obtained using a edge-based directional features of sizes 7 and 8 when KDA is utilized. In addition, window edge-based directional features for windows of size 8 and 9 have resulted in highest recognition rates for text-dependent Arabic text achieving 87.6 % for the top-ranked writer and 96.25 % for the top-ranked writers. For English handwriting, recognition rates of 75 % for the top-ranked writer and 90.75 % for the top-ranked writers using a window of size 5 were obtained when edge-based directional features are used for different text samples. In this case, a window of size 6 for edge-based directional feature produced the highest recognition rate for

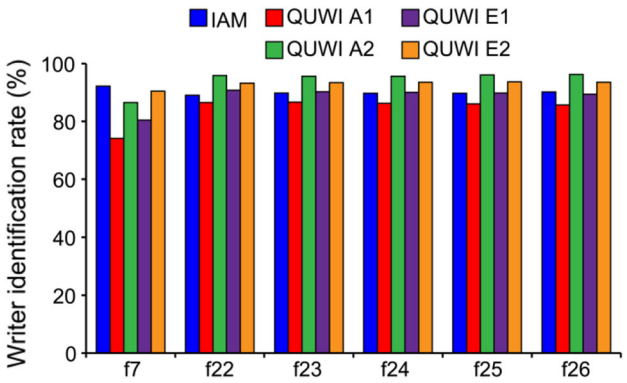

Fig. 7 Writer identification rate for the ten top-ranked writers using different datasets in Arabic and English

Table 3 State-of-the-art writer identification performance of individual features compared to the new features on the IAM dataset with 650 writers

Method	TOP1 (%)	TOP10 (%)
Contour-direction PDF [16]	46	76
Contour-hinge PDF [16]	81	92
Direction co-occurrence PDFs, horizontal run [16]	68	87
Direction co-occurrence PDFs, vertical run [16]	65	84
Grapheme emission PDF [16]	80	94
Distribution of chain codes [15]	36	74
Distribution of 1st-order differential chain codes [15]	34	76
Distribution of 2nd-order differential chain codes [15]	42	81
Distribution of chain code pairs [15]	67	88
Distribution of chain code triplets [15]	79	93
Distribution of curvature indices [15]	43	77
Local stroke direction distribution [15]	77	93
Distribution of 1st-order differential chain codes computed locally [15]	46	83
Distribution of 2nd-order differential chain codes computed locally [15]	42	79
Distribution of segment slopes [15]	55	86
Length-weighted distribution of segment slopes [15]	58	87
Distribution of curvatures [15]	37	75
Length-weighted distribution of curvatures [15]	40	78
Distribution of segment lengths [15]	31	72
Linear regression (f1)	46.70	76.70
Curvature (f2) using the x^2 distance	36.15	65.22
Tortuosity (f3)	47.50	79.10
Quadruple-order chain code(f7)	**82.70**	**92.20**
Edge-based directional features using the whole window size 2 (f18)	58.10	81.80
Edge-based directional features using the whole window size 3 f (19)	69.20	87.10
Edge-based directional features using the whole window size 4 (f20)	72.80	87.10
Edge-based directional features using the whole window size 5 (f21)	74.30	89.20
Edge-based directional features using the whole window size 6 (f22)	76.30	89.00
Edge-based directional features using the whole window size 7 (f23)	77.40	89.80
Edge-based directional features using the whole window size 8 (f24)	78.20	89.70
Edge-based directional features using the whole window size 9 (f25)	78.80	89.70
Edge-based directional features using the whole window size 10 (f26)	79.10	90.20

Bold values indicate the best results obtained

text-dependent handwritten text. For example, a recognition rate of 83.01 and 93.42 % was achieved for the top-ranked writers and 10 top-ranked writers, respectively.

Figure 7 shows the writer recognition rate for selected individual features (f7, f22–f26). These features have resulted in the highest recognition rate compared with other features. Features such as direction (f1), curvature (f2), and tortuosity (f3) can be used with other features to improve the recognition rate. Furthermore, features such as tortuosity (f3) can be used to measure the writing speed.

Table 3 shows the performance of the new features compared to the state-of-the-art counterparts when using the IAM dataset with approximately 650 writers. As shown in the table, the new features result in the highest recognition rate compared with the state-of-the-art features. The new fourth-order chain-code features show a significant improvement with regard to other features especially when combined with the KDA classifier.

5 Conclusion

This paper has presented several new features for writer identification. These include curvature, direction, and tortuosity features. We have shown through an evaluation and analysis of the recognition performance results obtained using two well-known datasets the usefulness of these features for writer identification. In particular, a comparison of the recognition performances of these features when deployed individually against state-of-the art features was also carried out. The findings have led us to propose methods to improve the discriminative power of both chain-code and edge-based directional features. To further ascertain the discriminative power of the proposed features we have utilized several distance metrics including X^2, L^2, and L^1 distances and the distance at the feature level when combined with KDA classifier by computing the differences between the feature vectors. The latter approach outperforms all the other distance measures previously reported in the literature. Work on the use of these features is ongoing for the prediction of demographic categories, including age, nationality, handedness, and gender. Future work will include the validation of the method with other languages and offline signature verification.

Acknowledgments This publication was made possible by a grant from the Qatar National Research Fund through the National Priority Research Program (NPRP) No. 09-864-1-128. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Qatar National Research Fund or Qatar University.

References

1. Srihari SN, Cha S-H, Arora H, Lee S (2001) Individuality of handwriting: a validation study. In: Sixth International Conference on Document analysis and recognition, 2001. Proceedings. IEEE, pp 106–109
2. Nosary A, Heutte L, Paquet T, Lecourtier Y (1999) Defining writer's invariants to adapt the recognition task. In: Proceedings of the Fifth International Conference on Document analysis and recognition, 1999. ICDAR'99. IEEE, pp 765–768
3. Franke K, Schomaker L, Koppen M (2005) Pen force emulating robotic writing device and its application. In: IEEE Workshop on Advanced robotics and its social impacts, 2005. IEEE, pp 36–46
4. Al-Ma'adeed S, Mohammed E, Al Kassis D (2008) Writer identification using edge-based directional probability distribution features for arabic words. In: IEEE/ACS International Conference on Computer systems and applications, 2008. AICCSA 2008. IEEE, pp 582–590
5. Marti U-V, Messerli R, Bunke H (2001) Writer identification using text line based features. In: Sixth International Conference on Document analysis and recognition, 2001. Proceedings. IEEE, pp 101–105
6. Franke K, Bunnemeyer O, Sy T (2002) Ink texture analysis for writer identification. In: Eighth International Workshop on Frontiers in Handwriting recognition, 2002. Proceedings. IEEE, pp 268–273
7. Said HE, Tan TN, Baker KD (2000) Personal identification based on handwriting. Pattern Recognit 33(1):149–160
8. Marti U-V, Bunke H (2002) The IAM-database: an English sentence database for offline handwriting recognition. Int J Doc Anal Recognit 5(1):39–46
9. Hertel C, Bunke H (2003) A set of novel features for writer identification. In: Audio-and video-based biometric person authentication. Springer, pp 679–687
10. Schlapbach A, Bunke H (2004) Using HMM based recognizers for writer identification and verification. In: Ninth International Workshop on Frontiers in Handwriting recognition, 2004. IWFHR-9 2004. IEEE, pp 167–172
11. Schlapbach A, Bunke H (2006) Off-line writer identification using Gaussian mixture models. In: 18th International Conference on Pattern Recognition, 2006. ICPR 2006. IEEE, pp 992–995
12. Bensefia A, Paquet T, Heutte L (2005) Handwritten document analysis for automatic writer recognition. Electron Lett Comput Vis Image Anal 5(2):72–86
13. Siddiqi IA, Vincent N (2007) Writer identification in handwritten documents. In: Ninth International Conference on Document analysis and recognition, 2007. ICDAR 2007. IEEE, pp 108–112
14. Siddiqi I, Vincent N (2009) A set of chain code based features for writer recognition. In: 10th International Conference on Document analysis and recognition, 2009. ICDAR'09. IEEE, pp 981–985
15. 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
16. Bulacu M, Schomaker L (2007) Text-independent writer identification and verification using textural and allographic features. IEEE Trans Pattern Anal 29(4):701–717
17. Imdad A, Bres S, Eglin V, Emptoz H, Rivero-Moreno C (2007) Writer identification using steered hermite features and svm. In: Ninth International Conference on Document analysis and recognition, 2007. ICDAR 2007. IEEE, pp 839–843
18. Muda A, Shamsuddin S, Darus M (2008) Invariants discretization for individuality representation in handwritten authorship. Comp Forensics pp 218–228
19. Bozeková M (2008) Comparison of handwritings. In: Central European Seminar on Computer Graphics (cescg. org)
20. Dolega B, Agam G, Argamon S (2008) Stroke frequency descriptors for handwriting-based writer identification. In: Electronic Imaging 2008. International Society for Optics and Photonics, pp 68150I–68158I
21. Steinke K-H, Gehrke M, Dzido R (2009) Writer recognition by combining local and global methods. In: 2nd International Congress on Image and Signal Processing, 2009. CISP'09. IEEE, pp 1–6
22. Jain R, Doermann D (2011) Offline Writer Identification using K-Adjacent Segments. In: 2011 International Conference on Document analysis and recognition (ICDAR). IEEE, pp 769–773
23. Maadeed SA, Ayouby W, Hassaine A, Aljaam JM (2012) QUWI: an Arabic and English handwriting dataset for offline writer identification. In: 2012 International Conference on Frontiers in Handwriting recognition (ICFHR). IEEE, pp 746–751
24. Hassaine A, Maadeed SA (2012) ICFHR 2012 Competition on writer identification challenge 2: Arabic Scripts. In: 2012 International Conference on Frontiers in handwriting recognition (ICFHR). IEEE, pp 835–840
25. Otsu N (1975) A threshold selection method from gray-level histograms. Automatica 11(285–296):23–27
26. Hassaïne A, Al-Maadeed S, Alja'am JM, Jaoua A, Bouridane A (2011) The ICDAR2011 Arabic writer identification contest. In: 2011 International Conference on Document analysis and recognition (ICDAR). IEEE, pp 1470–1474
27. Matas J, Shao Z, Kittler J (1995) Estimation of curvature and tangent direction by median filtered differencing. In: Image analysis and processing. Springer, pp 83–88
28. Zhang T, Suen CY (1984) A fast parallel algorithm for thinning digital patterns. Commun ACM 27(3):236–239
29. Koppenhaver K (2007) Forensic document examination: Principles and practice. Humana Press Inc., Totowa, New Jersey
30. Hassaïne A (2009) Restauration des pistes sonores optiques cinématographiques: approche par traitement d'images. Dissertation, École Nationale Supérieure des Mines de Paris