Writer identification using curvature-free features

He, Sheng; Schomaker, Lambertus

Published in:
Pattern recognition

DOI:
10.1016/j.patcog.2016.09.044

Publication date:
2017

Pattern Recognition 63 (2017) 451–464

Contents lists available at ScienceDirect

Pattern Recognition

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

Writer identification using curvature-free features

Sheng He*, Lambert Schomaker

Institute of Artificial Intelligence and Cognitive Engineering, University of Groningen, PO Box 407, 9700 AK Groningen, The Netherlands

ARTICLE INFO

Keywords:
Writer identification
Curvature-free
Run-lengths of local binary pattern
Cloud of line distribution

ABSTRACT

Feature engineering takes a very important role in writer identification which has been widely studied in the literature. Previous works have shown that the joint feature distribution of two properties can improve the performance. The joint feature distribution makes feature relationships explicit instead of roping that a trained classifier picks up a non-linear relation present in the data. In this paper, we propose two novel and curvature-free features: run-lengths of local binary pattern (LBPruns) and cloud of line distribution (COLD) features for writer identification. The LBPruns is the joint distribution of the traditional run-length and local binary pattern (LBP) methods, which computes the run-lengths of local binary patterns on both binarized and gray scale images. The COLD feature is the joint distribution of the relation between orientation and length of line segments obtained from writing contours in handwritten documents. Our proposed LBPruns and COLD are textural-based curvature-free features and capture the line information of handwritten texts instead of the curvature information. The combination of the LBPruns and COLD features provides a significant improvement on the CERUG data set, handwritten documents on which contain a large number of irregular-curvature strokes. The results of proposed features evaluated on other two widely used data sets (Firemaker and IAM) demonstrate promising results.

1. Introduction

Characterizing individual's handwriting style plays an important role in handwritten document analysis and automatic writer identification has attracted a large number of researchers in the pattern recognition field based on modern handwritten text [1], musical scores [2] and historical documents [3]. The writing patterns in handwritten documents encapsulate the individual's writing style in two aspects: the curvature of handwritten texts and the frequency of several basic patterns (graphemes), corresponding to the textural-based and grapheme-based algorithms. An observation can be found in the literature that the performance of textural-based methods is usually better than the performance of grapheme-based methods and combining them together often provides an improvement. In addition, the graphemes extracted from handwritten documents are easily visualized for end users. Therefore, both of them have been developed over the last decade.

Although the existing textural-based features have been successfully used for writer identification, many of them are not suitable for irregular-curvature handwriting, whose handwritten texts are often dominated by long straight-line segments, and polygonized, “hooked” corners, in writers with a low fluency. For example, as reported in [4], the performance (Top-1) of writer identification of Hinge [1] and Quill [5] are only 12.3% and 18.5% on the CERUG-EN data set, in which handwritings contain a large number of irregular-curvature strokes. The main reason is that Hinge and Quill feature methods focus on the fluent curvature of the ink trace and therefore exhibit a dramatic performance degradation on handwritten documents written by less skilled writers. The CERUG-EN data set contains handwritten texts in English written by Chinese subjects and it contains a large number of irregular-curvature strokes by two reasons: (1) Chinese writers tend to write line strokes affected by the habit of writing Chinese characters which are consisted of line-drawing strokes and (2) in real time, the velocity profile of on-line handwritings of non-native speakers shows pauses, as well as a degree of polygonization [6]. An example is shown in Fig. 1.

Previous works have shown that the probability distribution of the relation between two properties can improve the performance of writer identification. For example, the Hinge feature [1] is the probability distribution of orientations of two contour fragments attached at a common pixel. The Quill feature [5] is the probability distribution of the relation between the ink direction and the ink width and the oriented Basic Feature Columns (oBIF) [7] is the probability distribution of the bank of six Derivative-of-Gaussian filters on two scales. These features provide a significant improvement for writer identification.

* Corresponding author.
E-mail addresses: heshengxgd@gmail.com (S. He), L.Schomaker@ai.rug.nl (L. Schomaker).

http://dx.doi.org/10.1016/j.patcog.2016.09.044
Received 12 February 2016; Received in revised form 22 September 2016; Accepted 25 September 2016
Available online 28 September 2016

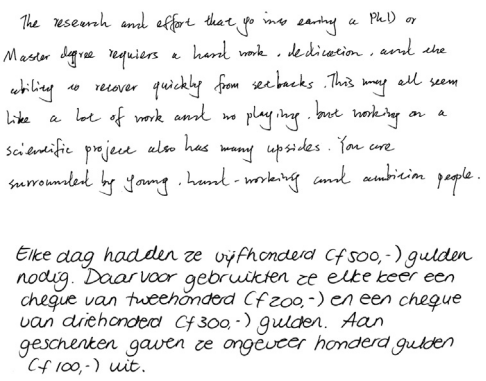

Fig. 1. The top figure shows an example of irregular-curvature strokes written by a non-native writer while the bottom figure shows fluent curvature strokes written by a native writer.

In this paper, we propose two curvature-free features for writer identification based on the run-lengths of general patterns, called run-lengths of local binary pattern (LBPruns) and the joint distribution of the relation between orientation and length of a set of line segments extracted on contours of ink traces, called cloud of line distribution (COLD). The traditional run-length method only considers one scanning line and only two simple patterns "0" and "1" are involved. Therefore, it fails to capture the spatial neighboring relationship between the simple patterns "0" and "1" over the neighbor lines of the scanning line. The proposed LBPruns can compute the run-lengths of more complex local binary patterns obtained by binary tests inspired by the LBP method [8]. Therefore, our proposed LBPruns can be considered as the general pattern run-length transform [9] which is a joint distribution of the traditional run-length and local binary pattern methods.

The writing contours can be approximated by a set of line segments using the polygon estimation method [10]. Generally, irregular-curvature handwritings with long ascenders and descenders lead to long lines in certain orientations while shaky and cursive strokes result in many short straight-lines in almost all directions [10]. We assume that the joint distribution of the relation between orientation and length of these straight-line segments can characterize the writing style. For example, the slopes of line segments reflect the slant information and the lengths of them reflect the curvature-based information (cursive handwritings lead to a large number of short lines and irregular-curvature handwritings result in a large number of long lines). The reference source codes will be available on the authors' website.

The rest of this paper is organized as follows. We summarize previous contributions to writer identification in Section 2 and we present our proposed LBPruns feature in Section 3 and the proposed COLD feature in Section 4. Section 5 provides the experimental results and Section 6 gives the conclusion.

2. Related work

Writer identification is the problem of recognizing the writer or author of a questioned document according to its handwriting style and it has been studied on different scripts, such as Arabic [11–14], English and Dutch [1,15], Chinese [4,16–19], Persian [20], Farsi [21] and Indic scripts [22–26]. A wide variety of features have been proposed for writer identification, which can be roughly divided into two groups: textural-based and grapheme-based features. Textural-based features are the statistical information about the slant and curvature of handwritten texts, while grapheme-based features extract local structures and map them into a common space, inspired by the bag-of-words model. A survey of writer identification before 1989 can be found in [27].

2.1. Textural-based feature

In the binarized image, the run-lengths of background pixels capture the properties of patterns enclosed spaces inside the characters and between letters and words. The probability distribution of run-lengths has been used for writer identification [28,29]. The gray level co-occurrence matrix (GLCM) [30,31], local binary patterns (LBP) [32,33] and local phase quantization (LPQ) [34] have been used to extract textural features based on texture blocks and have achieved promising results.

Filtering techniques have been studied to extract texture features from a handwritten text block for writer identification, such as the Gabor filter [35], XGabor filter [20] which is obtained by modulating a 2D centered sinusoid with a 2D Gaussian and oriented Basic Image Features (oBIF) Columns [7] which uses a bank of six Derivative-of-Gaussian filters to classify each location into seven possible symmetry types.

It has been shown that writing contours encapsulate the writing style of the writer and many features are extracted based on the handwritten contours. The joint distribution of the orientations of two legs of the obtained "hinge" based on edges [15] or contours [1] is used for writer identification and it has been extended to the Delta-n Hinge [36] to achieve the rotation-invariant property. The Quill feature [5], which is a probability distribution of the relation between the ink direction and the ink width, has been proposed for writer identification on both historical documents and modern handwriting to capture the property of writing instruments. Other types of contour based features, such as the distribution of chain codes and segment slopes, have also been studied in [10].

2.2. Grapheme-based feature

The COnnected-COmponent COntour (CO^3) has been proposed in [15] to isolate uppercase handwritten documents and it has been extended to lowercase handwriting with vector quantization based coding [1,37] and sparse coding [38] by segmenting cursive handwriting at the minima in the lower contour that are proximal to the upper contour (the detailed information can be found in [39]). Similar segmentation method has been proposed in [40] to build a pseudo-alphabet. The redundant patterns, which are the small parts of handwritten text without any semantic information, have been used in [10].

The typical bag-of-words model with the SIFT [41] feature has been used in [18,42,43] based on word regions. In [18], two types of features, SIFT Descriptor Signature (SDS) and Scale and Orientation Histogram (SOH), are extracted for writer identification. Similar works have been proposed in [44,45] with the Fisher kernel instead of the nearest neighbor coding. K-Adjacent Segments (KAS) features extracted on edges of documents are considered as the basic graphemes for writer identification [46] and script identification [47]. Zernike moments extracted on contours and encoded into Vectors of Locally Aggregated Descriptors (VLAD) has been proposed in [48] for writer identification.

Recently, the synthesized graphemes using the beta-elliptic model are used in [14] as the codebook for writer identification in Arabic handwriting, instead of obtaining the codebook from a training set. Singular structural regions in handwriting text, such as junctions, are extracted and considered as the basic graphemes in [4] for cross-script writer identification between Chinese and English.

3. Run-lengths of local binary pattern (LBPruns)

The "*run*" is defined as a sequence of connected pixels which have the same property (such as the gray value) in a given scanning line [29]. The lengths of these runs can be quantized into a histogram and the normalized histogram is considered as the run-length feature. For example, in a binary sequence "0001111010011" the run lengths of

value “0” are “3,1,2” and the run lengths of value “1” are “4,1,2”. The run-length feature is widely used in the document analysis community. It was first used for writer identification in [49] and on historical documents in [50]. In [51], the run-lengths histogram is used for document image retrieval and classification. Other applications of run-lengths can be found in [52,53].

However, the traditional run-length feature computes the run-lengths of the “0” and “1” based on one scanning line on binarized images and fails to capture the spatial correlation information of the run-lengths of these binary values with their neighbors. Although the correlation between two consecutive scanning lines has been used in [54,55] for text and non-text classification, the types of bit patterns (e.g., [0 0], [0 1], [1 0], [1 1]) are still limited.

In this section, we propose a general pattern run-length method based on several disparate scanning lines with certain inter-line distance between the consecutive scan-lines. For a position on several disparate scanning lines, the local binary pattern (LBP) code can be obtained directly from the scanning lines with binary values on binarized images or by thresholding their pixel values into binary values based on a reference scanning line on gray scale images. Then the run-lengths of the possible LBP codes are quantized into a histogram to form the feature representation.

3.1. LBPruns on binarized images (LBPruns_B)

Given n parallel scanning lines in certain direction (horizontal or vertical) with an inter-line distance d on a binarized image, the LBP code $\mathbf{p}$[1] on the position x is computed by:

$$\mathbf{p}(x) = \sum_{i=0}^{n-1} g_{y+i*d}(x)*2^i \tag{1}$$

where $g_{y+i*d}(x) \in \{0, 1\}$ is the binary pixel value on the position x of the scanning line $y + i*d$, y is the position of the first scanning line and * indicates the multiply between two integers. It is important to note that the LBP code $\mathbf{p}$ of the proposed LBPruns is obtained based on a translational symmetric neighbors instead of a circularly symmetric neighbors used in LBP [8]. In fact, there is also a binary test in Eq. (1), where the binary value $g_{y+i*d}(x)$ is obtained by a threshold involved in the processing of the image binarization.

Unlike the LBP method [8] which quantizes the LBP code into a histogram without considering the spatial relationship, we compute the histogram of the run-lengths of LBP code $\mathbf{p}$ in the same direction of the scanning line. In practice, we assume that the n scanning lines involved in the computation form a sequence of 2^n possible LBP codes $\mathcal{S}$. Given a certain LBP code $\mathbf{p}$, the sequence $\mathcal{S}$ can be converted into 0/1 string line $b_{\mathbf{p}}$ by:

$$b_{\mathbf{p}}(x) = \begin{cases} 1 & \text{if } \mathcal{S}(x) = \mathbf{p} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

The run lengths of the LBP code $\mathbf{p}$ in the sequence $\mathcal{S}$ can be obtained by counting the run-lengths of the value “1” in the converted string line $b_{\mathbf{p}}(x)$. Fig. 2 shows an example of the run-lengths with $n = 3$ scanning lines and the corresponding converted string lines of three LBP codes: (0,1,0), (0,1,1) and (1,1,1).

3.2. LBPruns on gray scale images (LBPruns_G)

In this section, we present a method to extract the run-lengths of LBP codes on gray scale images[2] without using any binarization method, inspired by LBP [8]. Given a center scanning line in a gray scale image, we find m “previous” scanning lines and m “succeeding” scanning lines with an inter-line distance d. We use l_y to denote the center scanning line and the set of other $2m$ scanning lines is denoted by $\mathcal{L} = \{l_{y-m*d}, l_{y-(m-1)*d}, \ldots, l_{y+(m-1)*d}, l_{y+m*d}\}$, where y denotes the position of the center line on the given image. The LBP code $\mathbf{p}$ on the position x of scanning lines is computed by:

$$\mathbf{p}(x) = \sum_{i=0}^{2m} s(g_y(x) - g_i(x), \theta)*2^i \tag{3}$$

$$s(x, \theta) = \begin{cases} 1 & \text{if } x < \theta \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where $g_y(x)$ and $g_i(x)$ are the pixel values on the position x of the center scanning line l_y and other scanning lines l_i in $\{y - m*d, y - (m - 1)*d, \ldots, y + (m - 1)*d, y + m*d)\}$, respectively. θ is the threshold for the binary test in Eq. (4). Fig. 3 illustrates a center scanning line with other four neighbors. Finally, the sequence $\mathcal{S}$ of 2^{2m} possible LBP codes can be converted into a binary string $b(x)$ given a certain LBP code $\mathbf{p}$, similar as Eq. (2). The run-lengths of the given LBP code $\mathbf{p}$ can be computed by counting the runs of the value “1” in the binary string $b(x)$.

Moreover, we can generalize the proposed method to compute the run-lengths of any given pattern. A binary test can be defined as:

$$b(x, \theta) = \begin{cases} 1 & \text{if } D(\mathcal{S}(x), \mathbf{p}) < \theta \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where $\mathbf{p}$ is the given pattern and $\mathcal{S}(x)$ is the element in the position x of the sequence $\mathcal{S}$, $D(\mathcal{S}(x), \mathbf{p})$ is a defined distance function and θ is a threshold. This method can convert the sequence $\mathcal{S}$ into a binary string given the pattern $\mathbf{p}$. Fig. 4 illustrates an example of the processing of converting a scanning line into a binary string. Then the run-lengths of the pattern $\mathbf{p}$ can be computed to be same as the ones of LBPruns_B and LBPruns_G. We will leave this method for future works.

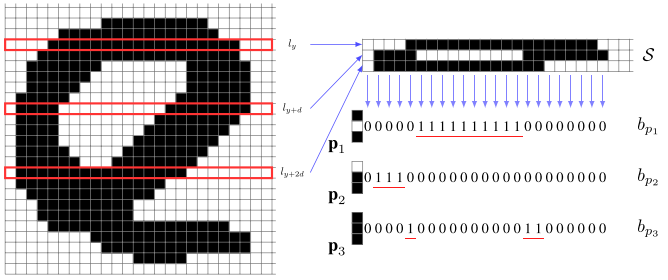

Fig. 2. The run-lengths of the more complex local binary pattern codes $\mathbf{p}_1 = (\mathbf{0}, \mathbf{1}, \mathbf{0})$, $\mathbf{p}_2 = (\mathbf{0}, \mathbf{1}, \mathbf{1})$, and $\mathbf{p}_3 = (\mathbf{1}, \mathbf{1}, \mathbf{1})$ on the sequence $\mathcal{S}$ formed by the three lines $(n = 3) l_y$, l_{y+d}, l_{y+2d} with the distance d=6.

3.3. LBPruns feature construction

We compute a run-length histogram of each LBP code $\mathbf{p}$ with a maximum length threshold $N_{max} = 100$ following the work [29] and this histogram is normalized. Finally, all the normalized histograms of all possible LBP codes are concatenated into one feature vector. Therefore, the feature dimensions are $2^n \times 100$ and $2^{2m} \times 100$ for LBPruns_B and LBPruns_G, respectively.

Our proposed method is different from LBP [8] in two aspects: (1) LBP computes the LBP codes in a circularly symmetric neighbors while the proposed method computes the LBP codes in a translational symmetric neighbors and (2) LBP computes the frequency of each LBP code while the proposed method considers the run-lengths of each LBP code, encoding the spatial information. In addition, our proposed method can be easily generalized to the run-lengths of arbitrary patterns (see Fig. 4).

[1] In fact, the $\mathbf{p}(x)$ in Eq. (1) is the number of LBP code, as defined in [8].
[2] In this paper, we assume that the pixel value on the gray scale images is in [0,255].

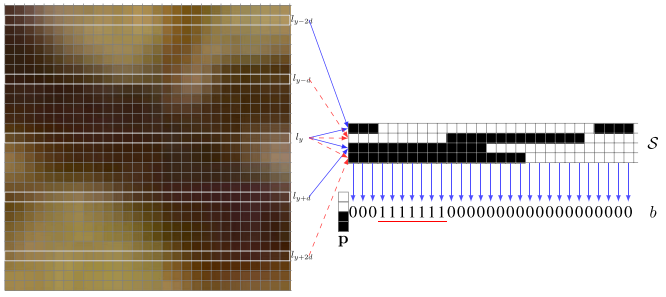

Fig. 3. The LBPruns_G computation in a gray scale image with d=6.

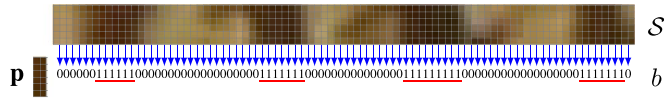

Fig. 4. The run-lengths of the arbitrary pattern **p** on the sequence S and b are the converted binary string.

4. COLD feature

The contours of connected components of handwritten texts contain the individual's handwriting style information, such as the writing slant and curvature [1]. Therefore, many researchers have taken efforts to extract features on contours to capture the curvature information. However, curvature-based methods fail on the irregular-curvature handwriting samples in which handwritten texts contain long straight lines. Therefore, in this section, we aim to design a novel curvature-free feature to capture writing styles of handwritten documents without considering the curvature information.

4.1. Pre-processing

The first step of the proposed method is to binarize the input handwritten document image. The Otsu thresholding [56] method, which is an efficient and parameterless global binarization method, is widely used on the clean modern handwriting images. In practice, we have found that there does not exist significant differences on binarized images obtained by Otsu, AdOtsu [57,58] and other binarization methods [59,60] on modern handwritten document images. Therefore, we adopt the Otsu threshold method in this paper.

After thresholding, the contours of connected components are extracted using the simple and robust method proposed in [5,15]. It starts at the left-most pixel of a connected component and traces the imaginary edges on the binarized image in a counterclockwise fashion, yielding a sequence of coordinates (x_i, y_i) of all of the edge pixels. Fig. 5(b) shows the extracted contour of the connected component of Fig. 5(a).

Based on the fact that every digital curve is composed of digital line segments [61], we decompose contours into maximal digital line segments by finding the dominant points on contours. This method is also known as polygonal approximation and is widely used in handwriting recognition [10,62] and shape classification [63,64]. In principle, any polygonal approximation approach can be applied to estimate the polygonal curve, such as the discrete contour evolution (DCE) [61]. Here, we use a parameter-free method proposed in [65] to detect the dominant points which are the vertices of the approximated polygonal curves. In order to remove the redundant dominant points, we adopt the constrained collinear-points suppression process proposed in [62]. Fig. 5(c) shows the detected dominant points (red points).

4.2. Cloud of line distribution

Given an ink contour S and the ordered sequence of n dominant points $\mathcal{P} = \{p_i(x_i, y_i),\ i = 0, 1, 2, \ldots, n\}$ from the contour, line segments can be obtained between every pair of the dominant points (p_i, p_{i+k}), where k denotes the distance on the dominant point sequence $\mathcal{P}$. Fig. 5(d) and (f) show line segments obtained with $k = 1$ and $k = 2$, respectively. The orientation θ and length ρ of each line segment can be measured by:

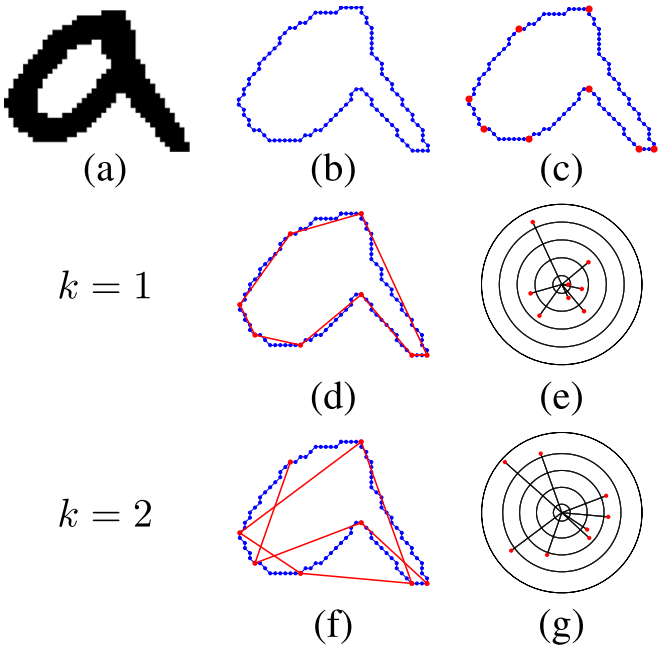

Fig. 5. Illustration of the process of the COLD construction: (a) the given binarized connected component; (b) the contour extracted from the binarized image (a); (c) detected dominant points (red points); (d) line segments (red lines) obtained between pair dominant points when $k = 1$; (e) the distribution of lines from (d) in the polar coordinate space; (f) line segments when $k = 2$ (note that some long lines are not shown in order to make the figure more clear); (g) the distribution of lines from (f) in the polar coordinate space. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

$$\begin{cases} \theta = \arctan\left(\dfrac{y_{i+k} - y_i}{x_{i+k} - x_i}\right) \\ \rho = \sqrt{(y_{i+k} - y_i)^2 + (x_{i+k} - x_i)^2} \end{cases} \tag{6}$$

where (x_i, y_i) and (x_{i+k}, y_{i+k}) are the coordinates of dominant point p_i and p_{i+k}, respectively. Each line corresponds to a point (θ, ρ) in the polar coordinate space (see Fig. 5(e) and (g)) and all line segments from one handwritten document can form a distribution, termed cloud of line distribution (COLD) . When $k = 1$, the line segments are the polygon estimation of the contours and the corresponding COLD reflects the slant and curvature-based information of contours. For example, in a more round handwriting, the lengths of line segments are short in all directions and the COLD has a high density around the origin. Note that dominant points are the high curvature points where the contour takes a turn. The straight-lines formed by the pair dominant points (p_i, p_{i+k}) where $k > 1$ indicate how long the pen moved in the Euclidean space when the contour turns $k - 1$ times, and the corresponding COLD can also capture some properties of the writing style.

Fig. 6 shows the COLDs of handwriting samples with $k = 1$ from three different writers, from which we can see that handwriting samples from the same hand have the similar line distributions and samples from different writers have different distributions. The differences of the COLDs are from the different densities (with different colors in Fig. 6) in different positions. Several important observations can be obtained from the COLDs in this figure. Firstly, densities in the regions closed to the center (the origin point) are high, which indicates

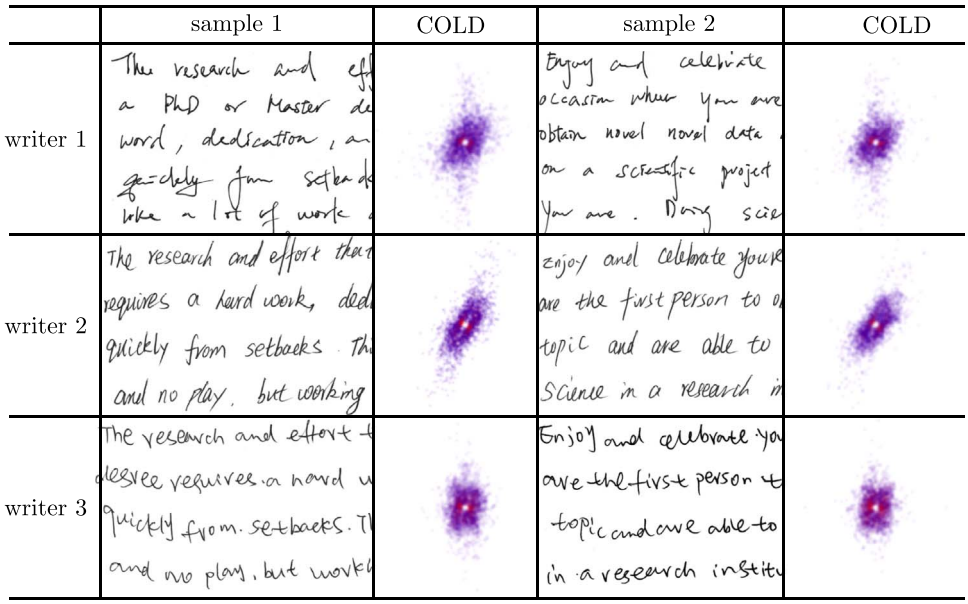

Fig. 6. Examples of COLDs of handwriting from three different subjects. The color closed to red in COLD means high density and the color closed to blue means low density. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

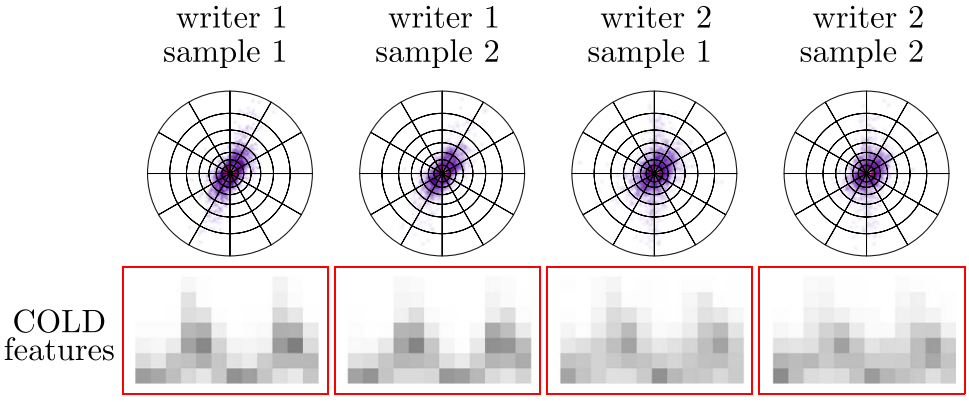

Fig. 7. COLD descriptors for handwriting samples from two writers. The top row shows the COLDs in the log-polar spaces. The bottom row shows the corresponding COLD features. The x-axis denotes the orientation bins and the y-axis denotes log distance bins.

Table 1
The best writer identification performance of the LBPruns on the CERUG data set with fixed parameters and the best performance found with the 10-fold cross-validation.

Feature	CERUG-CN		CERUG-EN		CERUG-MIXED	
	Top1	Top10	Top1	Top10	Top1	Top10
LBPrunsB$_{hv}$(5, 5)	88.6	95.7	77.1	98.1	90.9	100
LBPrunsG$_{hv}$(2,5,90)	86.7	95.7	88.6	99.0	88.1	99.5
LBPruns_B (10-fold)	89.2 ± 3.9	95.4 ± 2.3	86.1 ± 2.9	99.5 ± 0.6	94.2 ± 1.1	100 ± 0.0
LBPruns_G (10-fold)	87.1 ± 1.3	94.9 ± 1.5	93.4 ± 1.3	98.4 ± 1.2	92.7 ± 2.3	100 ± 0.0

that there are more short lines in handwritten documents. It is natural that many short lines are generated in order to estimate the high-curvature contours by the polygon shapes with a small error. Secondly, points in the regions far away from the center are sparse and the prevalent orientation corresponds to the slant of writing. Thirdly, the centralized COLD corresponds to the high curvature handwriting while the scattered COLD corresponds to the irregular-curvature handwriting.

From the above discussions we can see that the COLD reflects some attributes of handwriting and encapsulates the writing style of the corresponding handwritten document. Therefore, it can be used to build the feature descriptor to characterize the writing style.

4.3. Cold descriptor

Although the COLD captures the individual's writing style, it cannot be directly used as a feature descriptor. The main reason is that comparing the COLDs by a point-to-point way is sensitive to the variations between different handwriting samples from the same hand. Inspired by the Shape Context [66], we quantize the COLD into a log-polar histogram to compute the feature vector. The main advantage of using the log-polar space is that it makes the descriptor more sensitive

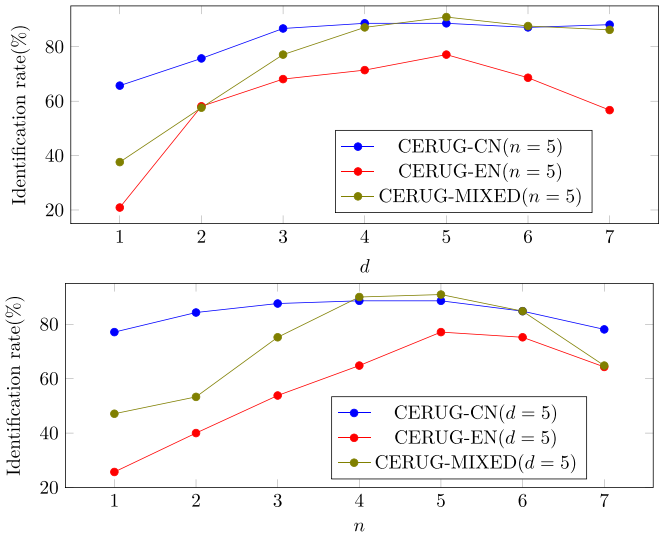

Fig. 8. The Top1 performance of the LBPruns_B feature with different parameters on the CERUG data set.

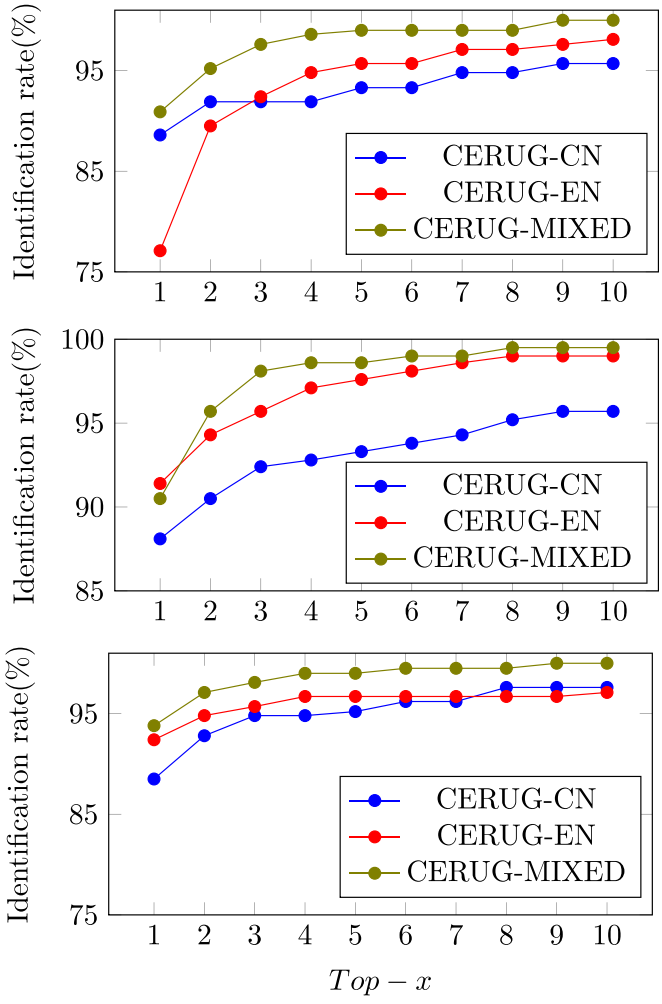

Fig. 9. The CMC curves of writer identification on the CERUG data set. The Top figure shows CMC curves of the $\text{LBPrunsB}_{hv}(5, 5)$ feature, the middle figure shows CMC curves of the $\text{LBPRunsG}_{hv}(2, 5, 90)$ feature and the bottom figure shows CMC curves of the COLD feature.

to regions of nearby the center than to those of regions farther away [66]. The normalized histogram is the final feature vector, which is the final COLD feature in this paper. Fig. 7 shows four COLDs in the log-polar space and their corresponding COLD features.

Table 2
Writer identification performance of the proposed COLD feature with different k on the CERUG data set.

COLD	Dimension	CERUG-CN		CERUG-EN		CERUG-MIXED	
		Top1	Top10	Top1	Top10	Top1	Top10
$k = 1$	84	89.0	97.1	80.9	95.2	74.7	99.0
$k = 2$	84	81.9	95.7	79.0	96.7	74.8	98.6
$k = 3$	84	71.9	92.9	81.4	97.1	65.2	95.7
$k = 4$	84	62.8	90.5	79.5	96.7	54.7	89.0
$k = 1, 2$	168	90.5	97.1	88.5	97.1	87.6	99.5
$k = 1, 2, 3$	252	88.5	97.6	92.4	97.1	93.8	100
$k = 1, 2, 3, 4$	336	88.6	96.7	92.4	97.6	92.4	100

There are three parameters involved in building the log-polar space. The distance between two consecutive rings in the log space D_c, the number of angular intervals N_p and the number of distance intervals N_q. In practice, we have found that the performance is stable when these parameters lie in certain ranges. In this paper, we empirically set them as: $D_c = 5$, $N_p = 12$ and $N_q = 7$. In addition, the COLD feature generated with a single k does not achieve the optimal performance, but a combination of COLDs with different k achieves the best performance. Therefore, we concatenate the COLDs with different k together to form the final feature vector.

5. Experiments

In this section, we use the proposed features to represent handwritten documents and the similarity between two writing samples is measured by the χ^2 distance. The nearest neighbor classification method is used for writer identification with a "leave-one-out" strategy. The query document is recognized as the writer of the document on the top x of the hit list, corresponding to the top-x performance, and the Top-1 and Top-10 results are reported.

We use $\text{LBPrunsB}_i(n, d)$ to denote the run-lengths of LBP feature computed on binarized images, where n is the number of scanning lines, d is the inter-line distance, and $i \in \{h, v, hv\}$ is the index of line directions and we only consider directions of horizontal (h), vertical (v) and the combination of horizontal and vertical (hv) directions. We use $\text{LBPrunsG}_i(m, d, \theta)$ to denote the run-lengths of LBP feature computed on gray scale images, where m is the number of the "previous" and "succeeding" scanning lines related to the center scanning line, d is the inter-line distance, θ is the threshold and i has the same meaning as it in the $\text{LBPrunsB}_i(n, d)$. The selection of these parameters is discussed on each data set.

5.1. Performance on the CERUG data set

The relatively new CERUG data set [4] contains handwritten documents written by 105 Chinese subjects in Chinese and English and each writer produced four pages (two pages in Chinese, one page in English and one page in both English and Chinese). We divided the CERUG data set into three subsets: CERUG-CN which contains Chinese handwriting, CERUG-EN which contains English handwriting and CERUG-MIXED which contains handwriting in both English and Chinese letters, following the method in [4]. As discussed in [4], the handwritten documents in the CERUG-EN data set have large straight lines and probability of line lengths greater than 100 is about 48 times and 8 times higher than the ones in Firemaker [67] and IAM [68]. Therefore, the CERUG-EN data set is considered as the curvature-free data set.

Table 3
The writer identification performance of the LBPruns and COLD features and their combinations on the CERUG data set.

Feature	CERUG-CN		CERUG-EN		CERUG-MIXED	
	Top1	Top10	Top1	Top10	Top1	Top10
LBPrunsB$_{hv}$(5,5)	88.6	95.7	77.1	98.1	90.9	100
LBPrunsG$_{hv}$(2,5,90)	86.7	95.7	88.6	99.0	88.1	99.5
COLD	88.5	97.6	92.4	97.1	93.8	100
COLD + LBPrunsB$_{hv}$(5,5)	93.3	96.2	95.2	98.1	98.5	100
COLD + LBPrunsG$_{hv}$(2,5,90)	93.8	96.7	96.2	98.1	97.1	100

Table 4
The writer identification performance of run-length based methods on the CERUG data set.

Feature	CERUG-CN		CERUG-EN		CERUG-MIXED	
	Top1	Top10	Top1	Top10	Top1	Top10
WRL$_h$ [49]	22.9	64.8	34.3	76.7	17.1	53.3
WRL$_v$ [49]	16.7	54.8	10.0	24.8	1.9	14.3
WRL$_{hv}$ [49]	35.2	77.1	22.4	37.1	7.6	25.7
IRL$_h$ [49]	52.4	82.4	61.9	90.5	72.8	93.8
IRL$_v$ [49]	47.6	82.4	10.4	23.8	64.8	93.8
IRL$_{hv}$ [49]	73.8	88.6	20.5	44.3	86.2	97.6
LBPrunsB$_h$(5,5)	81.9	93.8	87.1	98.5	84.3	99.5
LBPrunsB$_v$(5,5)	80.4	93.3	35.7	82.9	72.9	96.2
LBPrunsB$_{hv}$(5,5)	88.6	95.7	77.1	98.1	90.9	100
LBPrunsG$_h$(2,5,90)	80.5	91.4	86.7	98.5	73.8	97.6
LBPrunsG$_v$(2,5,90)	80.0	94.3	55.2	93.3	69.5	96.2
LBPrunsG$_{hv}$(2,5,90)	86.7	95.7	88.5	99.0	88.1	99.5

Table 5
The writer identification performance of different LBP-based features on the CERUG data set.

Feature	CERUG-CN		CERUG-EN		CERUG-MIXED	
	Top1	Top10	Top1	Top10	Top1	Top10
LBP [8]	44.8	68.1	11.9	26.7	70.9	91.9
LBPB_{hv}(5,5)	61.4	87.6	56.2	91.4	88.6	99.0
LBPG$_{hv}$(2,5,90)	51.9	80.9	50.0	88.6	80.9	98.6
LBPrunsB$_{hv}$(5,5)	88.6	95.7	77.1	98.1	90.9	100
LBPrunsG$_{hv}$(2,5,90)	86.7	95.7	88.5	99.0	88.1	99.5

Table 6
The performance of different line-based methods on the CERUG data set.

Feature	CERUG-CN		CERUG-EN		CERUG-MIXED	
	Top1	Top10	Top1	Top10	Top1	Top10
HOLD(k = 1)	11.4	53.3	9.0	46.2	15.2	50.0
HOSD(k = 1)	62.4	91.9	40.9	84.3	52.8	93.3
HOSD+HOAD(k = 1)	72.4	93.8	54.3	92.4	65.7	95.7
COLD(k = 1)	89.0	97.1	80.9	95.2	74.7	99.0
HOLD(k = 1, 2, 3)	34.3	70.5	29.0	80.9	41.9	81.4
HOSD(k = 1, 2, 3)	82.8	96.2	68.6	94.3	66.7	97.6
HOLD+HOSD(k = 1, 2, 3)	78.1	93.8	77.6	96.7	87.1	98.1
COLD(k = 1, 2, 3)	88.5	97.6	92.4	97.1	93.8	100

5.1.1. Parameter evaluation of LBPruns features

In this section, we evaluate the performance of writer identification on the CERUG data set with different parameters of LBPruns features by the 10-fold cross-evaluation. Each data set is randomly segmented into two approximately equal parts: one for the selection of the best parameters and another one for evaluation. The parameter spaces of n and d are from 1 to 7. We find the best value of m from 1 to 4 and the best threshold in $\theta \in \{60, 70, 80, 90, 100, 110, 120\}$. Finally, the average results with the standard deviations are reported in Table 1. Although we have found that the best results are obtained with different parameters on different data sets, we report the performance of LBPrunsB(5, 5) and LBPrunsG(2, 5, 90) on the three subsets of the CERUG data set in order to keep the parameter selection simple. In fact, from Table 1 we can see that the performance of LBPrunsB(5, 5) is not optimal on the CERUG-EN data set. Fig. 8 shows the Top-1 performance of LBPruns_B with different parameters. From the figure we can see that the number of scanning lines n is important, which determines the complexity of the LBP code. Similar trend is also found on the performance of the LBPruns_G feature. The top and middle figures in Fig. 9 show the cumulative match characteristic (CMC) curves [69] of the LBPruns_B and LBPruns_G features on each data set. The CMC curve plots the Top-x (x is from 1 to 10) performance of writer identification.

5.1.2. Parameter evaluation of the COLD feature

Table 2 shows the results of writer identification on the CERUG data set using the COLD feature and their combinations with different k. We can see that the performance decreases when k increases and the combined feature improves the identification rates. This observation is as expected, since combining COLD features with different k provides multi-scale information of writing contours. In the following experiments, we report the results of the COLD feature with k = 1, 2, 3 which provides reasonable results on the CERUG data set. The bottom figure in Fig. 9 shows the cumulative match characteristic (CMC) curves [69] of the COLD feature on each data set.

5.1.3. Performance of the combination of LBPruns and COLD features

In this section, we evaluate the performance of writer identification using the proposed LBPruns and COLD features. Since the LBPruns and COLD features capture different aspects of individual's writing style, combining them by distance averaging $d = \lambda d_{LBPruns} + (1 - \lambda) d_{COLD}$ improves, nevertheless, the performance, where λ is the coefficient. In all experiments in this paper, we set $\lambda = 0.1$ because the LBPruns feature is normalized based on the histogram of each LBP code and the sum of them is greater than 1, which means that the $d_{LBPruns}$ is greater than d_{COLD}. The value is based on experimental evaluation and the performance was maximal at $\lambda = 0.1$. Table 3 shows the performance of writer identification of the proposed individual features and feature combinations on the CERUG data set. From the table we can see that the recognition rates of LBPrunsB$_{hv}$(5, 5) and LBPrunsG$_{hv}$(2, 5, 90) obtained on three data sets

Table 7
The writer identification performance of different methods on the CERUG data set. Refer to Table 3 for individual COLD and LBPruns feature performance.

Feature	CERUG-CN		CERUG-EN		CERUG-MIXED	
	Top1	Top10	Top1	Top10	Top1	Top10
Hinge [1]	90.8	96.2	12.3	30.0	84.7	95.7
Quill [5]	82.7	92.3	15.8	48.6	74.8	93.3
Junclets [4]	90.4	**97.1**	87.1	96.2	85.7	98.5
COLD + LBPrunsB$_{hv}$(5,5)	93.3	96.2	95.2	**98.1**	**98.5**	**100**
COLD + LBPrunsG$_{hv}$(2,5,90)	**93.8**	96.7	**97.1**	**98.1**	97.1	**100**

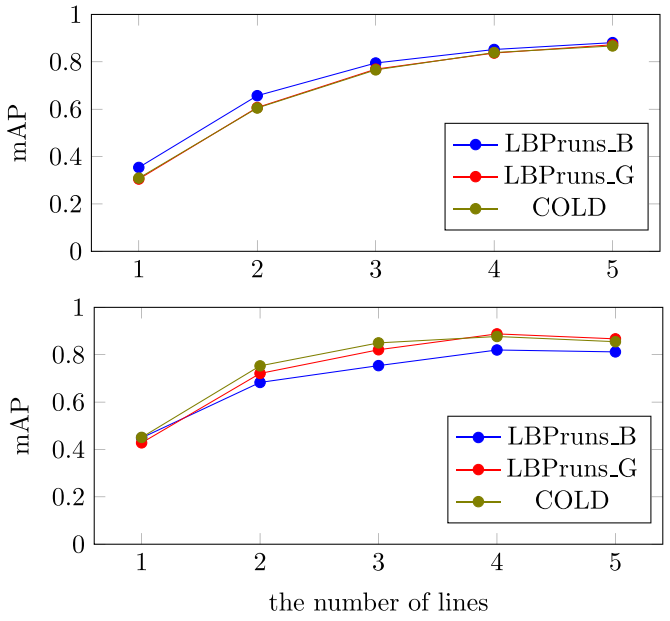

Fig. 10. The performance of writer retrieval of different features with different lines on the CERUG-CN (top figure) and CERUG-EN (bottom figure) data sets.

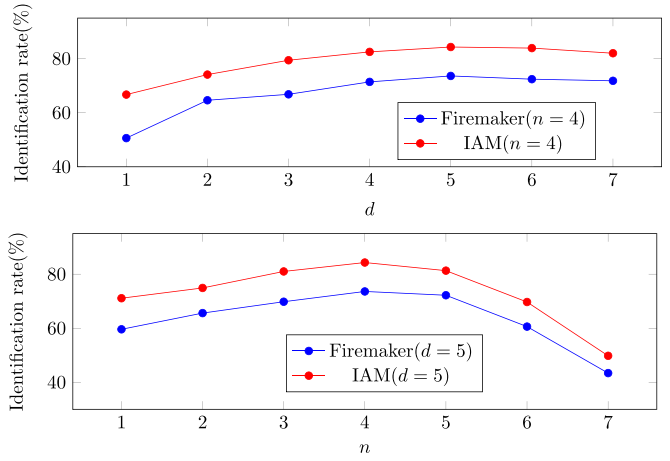

Fig. 11. The Top1 performance of the LBPruns_B feature with different parameters on the Firemaker and IAM data sets.

are very similar, except the Top-1 performance on the CERUG-EN data set. The performance of the COLD feature is slightly better than LBPruns features. It is important to note that combining LBPruns and COLD features produces significant improvements over the Top-1 performance and identification rates are 93.8%, 96.2% and 98.5% on the Chinese texts, English texts and mixed texts on the CERUG data set, respectively.

Table 8
The performance of writer identification of LBPruns on the Firemaker and IAM data sets with fixed parameters and the best performance found with the 10-fold cross-validation.

Feature	Firemaker		IAM	
	Top1	Top10	Top1	Top10
LBPrunsB$_{hv}$(4,5)	73.6	91.8	84.3	95.4
LBPrunsG$_{hv}$(2,5,90)	73.8	93.2	82.7	94.8
LBPruns_B (10-fold)	79.7 ± 3.0	95.8 ± 1.1	87.4 ± 1.4	96.4 ± 0.6
LBPruns_G (10-fold)	79.2 ± 2.5	96.9 ± 0.8	86.5 ± 2.2	96.4 ± 0.7

Table 9
The performance of writer identification of the proposed COLD feature with different k on the Firemaker and IAM data sets.

COLD with different k	Dimension	Firemaker		IAM	
		Top1	Top10	Top1	Top10
$k = 1$	84	77.4	92.0	75.5	91.5
$k = 2$	84	76.4	93.4	78.4	94.1
$k = 3$	84	72.6	93.0	72.3	92.5
$k = 4$	84	66.4	90.4	67.4	90.4
$k = 1, 2$	168	81.8	93.6	83.3	94.9
$k = 1, 2, 3$	252	**83.0**	94.6	83.6	**95.9**
$k = 1, 2, 3, 4$	336	79.8	**95.4**	**83.8**	95.6

5.1.4. Comparison with other studies

Table 4 shows the performance of traditional run-lengths of white pixel WRL$_i$ and ink pixel IRL$_i$ in horizontal and vertical directions and their feature combinations. We can see that the run-lengths of LBP codes perform much better than the run-lengths of "0" and "1". The benefits are from two aspects: (1) the LBP code can depict more complex patterns than "0" and "1" and (2) the supporting region of n or $2m$ scanning lines is larger than the single line. Therefore, the LBPruns features are more discriminative than the traditional run-length methods.

We also compare the LBPruns with the traditional LBP-based features. For LBP histogram, we follow the work [33] to keep 255 bins and the binary test is performed in a 3-by-3 neighborhood of each pixel. In addition, we compute the histogram of the LBP codes obtained from the n scanning lines on binarized images, denoted as LBP_B and from the $2m$ scanning lines on gray scale images, denoted as LBP_G, instead of computing the histogram of the run-lengths of the LBP codes. The difference between LBP and LBP_B (or LBP_G) is that LBP computes the LBP codes on a circularly symmetric neighbors while LBP_B (or LBP_G) computes the LBP codes on several parallel scanning lines in a certain direction. For fair comparison, we use the same parameters of LBPruns_B and LBPruns_G for LBP_B and

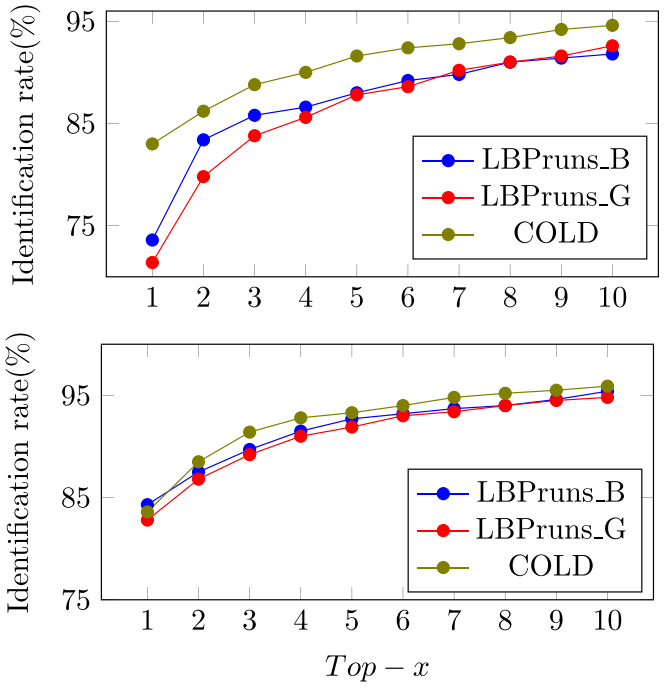

Fig. 12. The CMC curves of different features on the Firemaker (top figure) and IAM (bottom figure) data sets.

Table 10
The performance of writer identification of the LBPruns and COLD features and their combinations on the Firemaker and IAM data sets.

Feature	Firemaker		IAM	
	Top1	Top10	Top1	Top10
LBPrunsB$_{hv}$(4,5)	73.6	91.8	84.3	95.4
LBPrunsG$_{hv}$(2,5,90)	73.8	93.2	82.7	94.8
COLD	83.0	94.6	83.6	95.9
COLD + LBPrunsB$_{hv}$(4,5)	86.2	96.6	89.9	96.9
COLD + LBPrunsG$_{hv}$(2,5,90)	85.4	96.6	89.5	97.2

Table 11
The performance of writer identification of run-length based methods on the Firemaker and IAM data sets.

Feature	Firemaker		IAM	
	Top1	Top10	Top1	Top10
WRL$_h$ [49]	21.6	55.2	13.7	36.5
WRL$_v$ [49]	17.0	51.2	13.9	36.5
WRL$_{hv}$ [49]	40.8	76.6	31.4	58.0
IRL$_h$ [49]	22.8	46.6	37.6	68.1
IRL$_v$ [49]	31.0	59.6	54.8	81.2
IRL$_{hv}$ [49]	44.0	66.4	71.2	89.0
LBPrunsB$_h$(4,5)	68.2	89.4	81.2	93.6
LBPrunsB$_v$(4,5)	68.6	89.6	72.4	89.8
LBPrunsB$_{hv}$(4,5)	73.6	91.8	84.3	95.4
LBPrunsG$_h$(2,5,90)	63.4	87.4	72.8	91.7
LBPrunsG$_v$(2,5,90)	64.0	89.8	72.4	91.0
LBPrunsG$_{hv}$(2,5,90)	73.8	93.2	82.7	94.8

Table 12
The performance of writer identification of different LBP features on the Firemaker and IAM data sets.

Feature	Firemaker		IAM	
	Top1	Top10	Top1	Top10
LBP [8]	51.2	80.2	62.8	83.5
LBPB$_{hv}$(4,5)	48.8	78.0	64.5	87.9
LBPG$_{hv}$(2,5,90)	51.4	80.0	61.3	86.6
LBPrunsB$_{hv}$(4,5)	73.6	91.8	84.3	95.4
LBPrunsG$_{hv}$(2,5,90)	73.8	93.2	82.7	94.8

Table 13
The performance of writer identification of different line-based methods on the Firemaker and IAM data sets.

Feature	Firemaker		IAM	
	Top1	Top10	Top1	Top10
HOLD(k = 1)	21.4	61.0	13.9	47.0
HOSD(k = 1)	39.6	80.4	39.2	72.5
HOSD+HOAD(k = 1)	64.6	89.6	59.8	87.8
COLD(k = 1)	77.4	92.0	75.5	91.5
HOLD (k = 1, 2, 3)	47.4	77.4	44.8	73.2
HOSD(k = 1, 2, 3)	63.8	87.2	64.7	86.5
HOLD+HOSD(k = 1, 2, 3)	74.2	91.4	77.5	94.2
COLD(k = 1, 2, 3)	83.0	94.6	83.6	95.9

Table 14
The performance of writer identification of different features on the Firemaker and IAM data sets. Refer to Table 10 for individual COLD and LBPruns feature performance.

Feature	Firemaker		IAM	
	Top1	Top10	Top1	Top10
Hinge [1]	85.8	95.8	86.6	95.2
Quill [5]	60.8	78.8	84.6	93.8
Junclets [4]	80.6	94.0	83.3	94.4
COLD + LBPrunsB$_{hv}$(4,5)	**86.2**	**96.6**	**89.9**	96.9
COLD + LBPrunsG$_{hv}$(2,5,90)	85.4	**96.6**	89.5	**97.2**

Table 15
The performance of writer identification of different approaches on the Firemaker and IAM data sets.

Approach	Firemaker			IAM		
	Writers	Top-1	Top-10	Writers	Top-1	Top-10
Wu et al. [18]	250	92.4	98.9	657	98.5	99.5
Siddiqi and Vincent [10]	–	–	–	650	91	97
Bulacu and Schomaker [1]	250	83	95	650	89	97
Ghiasi and Safabakhsh [72]	250	89.2	98.6	650	93.7	97.7
Jain and Doermann [46]	–	–	–	300	93.3	96.0
He and Schomaker [36]	250	90.4	98.2	650	93.2	97.2
He and Schomaker [4]	250	89.8	96.0	650	91.1	97.2
Proposed	250	86.2	96.6	650	89.9	96.9

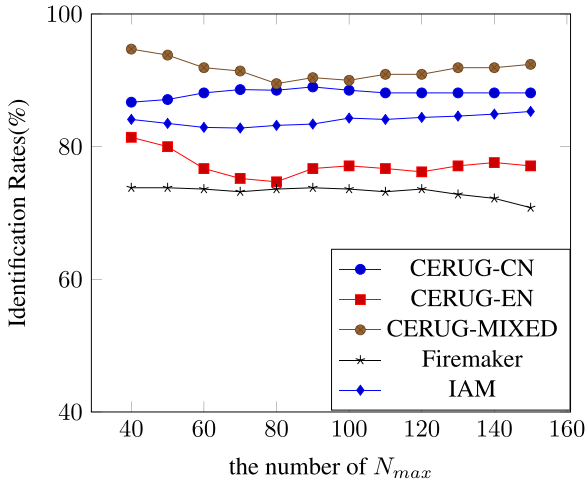

Fig. 13. The Top-1 performance of the LBPrunsB$_{hv}$ for the maximal run-length parameter N_{max}, showing stability of performance for this parameter on different data sets.

LBP_G. Table 5 shows the performance of writer identification using different LBP-based methods. From the table we can observe that the performance of the run-lengths of LBP codes exceeds the LBP, LBP_B and LBP_G features. The reason is that LBPruns computes the run-lengths of LBP codes which encodes the spatial information of these LBP codes and therefore can increase the discriminativeness of the features.

The slope and length distributions of line segments have also been used for writer identification in [10], which computes two histograms of slope and length distributions, separately. In order to demonstrate the powerful of our proposed COLD feature, we also compare it with the histogram of slope distribution (HOSD) and the histogram of length distribution (HOLD) and their linear combinations. The parameters are set the same as the COLD feature for fair comparison. Table 6 shows the results on the CERUG data set, which shows that our proposed COLD feature outperforms all the other features. It is also important to note that combining the line distributions with different k improves the performance of both the HOSD and HOLD features, as well as the proposed COLD feature. The reasons are that: (1) the COLD feature captures the joint distribution of slope and length distributions of line segments; (2) the COLD feature considers line distributions in a large scale when $k > 1$ while the method in [10] only considers line distributions with $k = 1$. In fact, the HOSD and HOLD can be considered as the marginal integrations of the COLD feature along slope and length directions, respectively.

We compare the proposed methods with several existing methods in the literature on the CERUG data set and experimental results are presented in Table 7. It is important to note that the curvature-based methods, such as Hinge [1] and Quill [5], fail on the curvature-free CERUG-EN data set and the Top-1 performance of Hinge and Quill are only 12.3% and 15.8%. The combination of the COLD and LBPruns features significantly improves the performance on the CERUG-EN data set.

5.1.5. Writer retrieval

The whole handwritten documents are separated into different number of text lines by the text detection method proposed in [70]. This means that each writer has more than two samples. Therefore, we perform the writer retrieval to evaluate the performance of the proposed features, using the measurement of the mean Average Precision (mAP). Fig. 10 shows the performance of writer retrieval of the proposed features on the CERUG-CN and CERUG-EN data sets, with the number of text lines from 1 to 5. From the figure we can see that the results of different features are similar on these two data sets. In addition, the performance is relatively more stable on handwritten documents with at least three lines (also found in [7,10]), containing approximately 100 characters, which is the minimum amount of needed text for writer identification using textural-based features [71].

5.2. Performance on the cursive data sets

We also evaluate the proposed curvature-free features on two widely used data sets: the Firemaker [67] and the IAM [68] data sets.

There are 250 writers on the Firemaker data set, where each writer produced four pages. We perform writer identification of page 1 versus 4, which were written using lowercase characters. We modified the IAM data set to make sure that each writer has two samples following the method in [1,10]: the first two handwritten images of writers who produced at least two pages are kept and images of writers who only contributed one page are divided into two parts. Finally, there are 650 writers on the modified IAM data set.

5.2.1. Parameter selection

Fig. 11 shows the Top-1 performance of LBPrunsB$_{hv}$ on the two data sets with different parameters of the LBPruns feature. From the figure we can see that the best results are achieved when $n = 4$ and $d = 5$. In practice, we have found that the LBPrunsG$_{hv}$ performs well with $m = 2$, $d = 5$ and $\theta = 90$ on the Firemaker and IAM data sets. Therefore, we report the results of LBPrunsB$_{hv}(4, 5)$ and LBPrunsG$_{hv}(2, 5, 90)$ on the Firemaker and IAM data sets in the following experiments. We also conduct the 10-fold cross-evaluation on the Firemaker and IAM data sets, and the performance is shown in Table 8.

Table 9 shows the performance of the COLD feature with different k on the two data sets. There is no obvious difference between the performance of COLD features with the combination of $k = 1, 2, 3$ and $k = 1, 2, 3, 4$, except that the top-1 performance of the COLD feature on Firemaker with $k = 1, 2, 3, 4$ is low. Therefore, we report the performance of the COLD feature with $k = 1, 2, 3$. Fig. 12 shows the CMC curves of the proposed features on the Firemaker and IAM data sets.

5.2.2. Performance of LBPruns and COLD features

Table 10 shows the results of the proposed LBPruns and COLD features and their combinations on the Firemaker and IAM data sets. We can see that the performance of the COLD feature is better than LBPruns features on the Firemaker data set and is comparable to the LBPruns on the IAM data set. Combining the LBPruns and COLD features outperforms all individual features involved in the combination.

5.2.3. Comparison with other studies

We also compare the proposed LBPruns with the traditional run-length and LBP methods on the Firemaker and IAM data sets, as the same experimental setting of the CERUG data set. Table 11 shows the results of the traditional white and ink run-length methods and the proposed LBPruns features, from which we can see that the proposed LBPruns methods consistently outperform the traditional ones. Table 12 shows the performance of LBPruns compared with the traditional LBP based methods and we can see that the run-lengths of LBP show superior performance with significant margin to the traditional LBP based methods. Table 13 shows the performance of the proposed COLD feature comparing to the traditional HOSD, HOLD and their combinations. From the table we can see that our proposed COLD feature gives significant improvements on the Firemaker and IAM data sets. Table 14 shows that the combination of the LBPruns and COLD features achieves the best results on Firemaker and IAM, comparing to the curvature-based Hinge and Quill features and the grapheme-based Junclets feature. Table 15 summarizes results of several works in the literature of writer identification on the

Fig. 14. Samples of different books in the Monk [73] system and their corresponding COLDs.

Firemaker and IAM data sets. Although our methods do not give state-of-the-art results on the cursive data sets, the LBPruns and COLD provide good results on the curvature-less CERUG data set.

5.3. The effect of the parameter N_{max} of the LBPruns method

In this experiment, we perform writer identification using LBPrunsB_{hv} with different maximum length threshold N_{max} on the CERUG, Firemaker and IAM data sets. The Top-1 performance is shown in Fig. 13. From the figure we can see that results are quite stable when $N_{max} \in [40, 150]$. As mentioned above, the $N_{max} = 100$ is used in this paper.

5.4. The COLD feature on other images

Our proposed COLD feature can also be used to capture the line structures on both historical documents and natural images. Fig. 14 shows samples of historical documents from 12 books in the Monk system [73] and their corresponding COLDs. From the figure we can see that the COLDs are quite different for documents from different books. For example, the fifth and sixth documents in the first row exhibit a strong slant in the diagonal direction and the Chinese woodblock printed document (the last one in the second row) shows long lines in the horizontal and vertical directions..

We can also apply our proposed COLD on natural images. We use the fast line detection method (LSD) proposed in [75] to detect the straight lines on natural images and use the extracted line segments to build the corresponding COLDs. Fig. 15 shows the corresponding COLDs on images from the demo of the paper [74]. Generally, the indoor images exhibit a strong structure and contain more long lines in a number of limited directions. However, scene images have a high textural information and contain more short lines in all directions and their COLDs are more centralized. We evaluate our proposed COLD feature with the spatial pyramid method [76] on the fifteen scene categories data set [77] using the k nearest neighbor classification and the recognition rate we achieved is around 44% when $k \in [10, 50]$, with a single feature, much less elaborate set up than [77].

Fig. 15. The first and fourth columns show that images are from demo of the paper [74]. The second and fifth columns show the lines extracted by LSD [75] method, and the third and sixth columns show the corresponding COLDs.

6. Conclusion

In this work we have introduced two novel curvature-free features: the run-lengths of Local Binary Pattern (LBPruns) which is the run-lengths histogram of local binary patterns and can be used on binarized images and gray scale images, and the cloud of line distribution (COLD) which is the distribution of line segments from contours of handwritten texts in the polar coordinate space and it is quantized into a log-polar histogram.

From the experimental results of writer identification on the CERUG, Firemaker and IAM data sets, we can conclude that our proposed LBPruns and COLD features work much better on the CERUG data set and the performance of their combination is comparable to other traditional features on the Firemaker and IAM data sets. In addition, the LBPruns method is the combination of traditional run-lengths and LBP methods and achieves much better results than run-lengths and LBP methods. We have explained the possible reasons in the previous sections that (1) LBPruns computes the run-lengths of more complex patterns than the simple “0” and “1” and hence it is more discriminative than the traditional run-length methods; (2) LBPruns computes the histogram of run-lengths of local binary pattern instead of the histogram of local binary pattern, thus it encodes the spatial information. The number of scanning lines involved in the LBPruns determines the complexity of the LBP codes and the inter-line distance reflects the scale information.

Furthermore, we have visualized the COLDs on both historical documents and natural images. We have shown that the COLD can capture the line structures on images which can be used, in future, for historical document retrieval and scene classification.

Acknowledgments

This work has been supported by the Dutch Organization for Scientific Research NWO (project No. 380-50-006). The authors would like to thank Shijie Zhao and Yanfang Feng to collect the CERUG data set.

References

[1] M. Bulacu, L. Schomaker, Text-independent writer identification and verification using textural and allographic features, IEEE Trans. Pattern Anal. Mach. Intell. 29 (4) (2007) 701–717.
[2] A. Gordo, A. Fornés, E. Valveny, Writer identification in handwritten musical scores with bags of notes, Pattern Recognit. 46 (5) (2013) 1337–1345.
[3] D. Arabadjis, F. Giannopoulos, C. Papaodysseus, S. Zannos, P. Rousopoulos, M. Panagopoulos, C. Blackwell, New mathematical and algorithmic schemes for pattern classification with application to the identification of writers of important ancient documents, Pattern Recognit. 46 (8) (2013) 2278–2296.
[4] S. He, M. Wiering, L. Schomaker, Junction detection in handwritten documents and its application to writer identification, Pattern Recognit. 48 (12) (2015) 4036–4048.
[5] A. Brink, J. Smit, M. Bulacu, L. Schomaker, Writer identification using directional ink-trace width measurements, Pattern Recognit. 45 (1) (2012) 162–171.
[6] R.G. Meulenbroek, G.P. Van Galen, The acquisition of skilled handwriting: discontinuous trends in kinematic variables, Adv. Psychol. 55 (1988) 273–281.
[7] A.J. Newell, L.D. Griffin, Writer identification using oriented Basic Image Features and the Delta encoding, Pattern Recognit. 47 (6) (2014) 2255–2265.
[8] T. Ojala, M. Pietikäinen, T. Mäenpää, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002) 971–987.
[9] S. He, L. Schomaker, General pattern run-length transform for writer identification, in: International Workshop on Document Analysis Systems, 2016, pp. 60–65.
[10] I. Siddiqi, N. Vincent, Text independent writer recognition using redundant writing patterns with contour-based orientation and curvature features, Pattern Recognit. 43 (11) (2010) 3853–3865.
[11] S. Gazzah, N. E. Ben Amara, Arabic handwriting texture analysis for writer identification using the DWT-lifting scheme, in: International Conference on Document Analysis and Recognition, vol. 2, 2007, pp. 1133–1137.
[12] M. Bulacu, L. Schomaker, A. Brink, Text-independent writer identification and verification on offline Arabic handwriting, in: International Conference on Document Analysis and Recognition, 2007, pp. 769–773.
[13] S. N. Srihari, G. R. Ball, Writer verification of Arabic handwriting, in: International Workshop on Document Analysis Systems, 2008, pp. 28–34.
[14] M.N. Abdi, M. Khemakhem, A model-based approach to offline text-independent Arabic writer identification and verification, Pattern Recognit. 48 (5) (2015) 1890–1903.
[15] L. Schomaker, M. Bulacu, Automatic writer identification using connected-component contours and edge-based features of uppercase western script, IEEE Trans. Pattern Anal. Mach. Intell. 26 (6) (2004) 787–798.
[16] Z. He, X. You, Y.Y. Tang, Writer identification of Chinese handwriting documents using hidden Markov tree model, Pattern Recognit. 41 (4) (2008) 1295–1307.
[17] J. Wen, B. Fang, J. Chen, Y. Tang, H. Chen, Fragmented edge structure coding for Chinese writer identification, Neurocomputing 86 (2012) 45–51.
[18] X. Wu, Y. Tang, W. Bu, Offline text-independent writer identification based on scale invariant feature transform, IEEE Trans. Inf. Forensics Secur. 9 (3) (2014) 526–536.
[19] W. Yang, L. Jin, M. Liu, Chinese character-level writer identification using path signature feature, dropstroke and deep CNN, in: International Conference on Document Analysis and Recognition, 2015, pp. 546–550.
[20] B. Helli, M.E. Moghaddam, A text-independent Persian writer identification based on feature relation graph (FRG), Pattern Recognit. 43 (6) (2010) 2199–2209.
[21] F. S. Nejad, M. Rahmati, A new method for writer identification and verification based on Farsi/Arabic handwritten texts, in: International Conference on Document Analysis and Recognition, vol. 2, 2007, pp. 829–833.
[22] U. Garain, T. Paquet, Off-line multi-script writer identification using ar coefficients, in: International Conference on Document Analysis and Recognition, 2009, pp. 991–995.
[23] K. Karunakara, B. Mallikarjunaswamy, Writer identification based on offline handwritten document images in Kannada language using empirical mode decomposition method, Int. J. Comput. Appl. 30 (6) (2011) 31–36.
[24] S. Biswas, A. K. Das, Writer identification of Bangla handwritings by radon transform projection profile, in: International Workshop on Document Analysis Systems, 2012, pp. 215–219.
[25] S. Chanda, K. Franke, U. Pal, Text independent writer identification for Oriya script, in: International Workshop on Document Analysis Systems, 2012, pp. 369–373.
[26] C. Adak, B. Chaudhuri, Writer identification from offline isolated Bangla characters and numerals, in: International Conference on Document Analysis and Recognition, 2015, pp. 486–490.
[27] R. Plamondon, G. Lorette, Automatic signature verification and writer identification the state of the art, Pattern Recognit. 22 (2) (1989) 107–131.
[28] B. Arazi, Handwriting Identification by Means of Run-Length Measurements, IEEE Trans. Syst. Man Cybern. 7 (1977) 878–881.
[29] C. Djeddi, I. Siddiqi, L. Souici-Meslati, A. Ennaji, Text-independent writer recognition using multi-script handwritten texts, Pattern Recognit. Lett. 34 (10) (2013) 1196–1202.
[30] K. Franke, O. Bünnemeyer, T. Sy, Ink texture analysis for writer identification, in: International Workshop on Frontiers in Handwriting Recognition, 2002, pp. 268–273.
[31] R. Hanusiak, L.S. Oliveira, E. Justino, R. Sabourin, Writer verification using texture-based features, Int. J. Doc. Anal. Recognit. 15 (3) (2012) 213–226.
[32] A. Nicolaou, A. D. Bagdanov, M. Liwicki, D. Karatzas, Sparse radial sampling lbp for writer identification, in: International Conference on Document Analysis and Recognition, 2015, pp. 716–720.
[33] Y. Hannad, I. Siddiqi, M.E.Y. El Kettani, Writer identification using texture descriptors of handwritten fragments, Expert Syst. Appl. 47 (2016) 14–22.
[34] D. Bertolini, L.S. Oliveira, E. Justino, R. Sabourin, Texture-based descriptors for writer identification and verification, Expert Syst. Appl. 40 (6) (2013) 2069–2080.
[35] H.E. Said, T.N. Tan, K.D. Baker, Personal identification based on handwriting, Pattern Recognit. 33 (1) (2000) 149–160.
[36] S. He, L. Schomaker, Delta-n hinge: rotation-invariant features for writer identification, in: International Conference on Pattern Recognition, 2014, pp. 2023–2028.
[37] E. Khalifa, S. Al-maadeed, M. Tahir, A. Bouridane, A. Jamshed, Off-line writer identification using an ensemble of grapheme codebook features, Pattern Recognit. Lett. 59 (2015) 18–25.
[38] R. Kumar, B. Chanda, J. Sharma, A novel sparse model based forensic writer identification, Pattern Recognit. Lett. 35 (2014) 105–112.
[39] L. Schomaker, K. Franke, M. Bulacu, Using codebooks of fragmented connected-component contours in forensic and historic writer identification, Pattern Recognit. Lett. 28 (6) (2007) 719–727.
[40] R. Jain, D. Doermann, Writer identification using an alphabet of contour gradient descriptors, in: International Conference on Document Analysis and Recognition, 2013, pp. 550–554.
[41] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.
[42] S. Fiel, R. Sablatnig, Writer retrieval and writer identification using local features, in: International Workshop on Document Analysis Systems, 2012, pp. 145–149.
[43] Y. Xiong, Y. Wen, S. Wang, Y. Lu, Text-independent writer identification using SIFT descriptor and contour-directional feature, in: International Conference on Document Analysis and Recognition, 2015, pp. 91–95.
[44] S. Fiel, R. Sablatnig, Writer identification and writer retrieval using the fisher vector on visual vocabularies, in: International Conference on Document Analysis and Recognition, 2013, pp. 545–549.
[45] R. Jain, D. Doermann, Combining local features for offline writer identification, in: International Conference on Frontiers in Handwriting Recognition, 2014.
[46] R. Jain, D. Doermann, Offline writer identification using k-adjacent segments, in: International Conference on Document Analysis and Recognition, 2011, pp. 769–773.
[47] G. Zhu, X. Yu, Y. Li, D. Doermann, Language identification for handwritten document images using a shape codebook, Pattern Recognit. 42 (12) (2009) 3184–3191.
[48] V. Christlein, D. Bernecker, E. Angelopoulou, Writer identification using VLAD encoded contour-Zernike moments, in: International Conference on Document Analysis and Recognition, 2015, pp. 906–910.
[49] B. Arazi, Handwriting identification by means of run-length measurements, IEEE Trans. Syst. Man Cybern. 12 (1977) 878–881.
[50] I. Dinstein, Y. Shapira, Ancient hebraic handwriting identification with run-length histograms, IEEE Trans. Syst. Man Cybern 12 (3) (1982) 405–409.
[51] A. Gordo, F. Perronnin, E. Valveny, Large-scale document image retrieval and classification with runlength histograms and binary embeddings, Pattern Recognit. 46 (7) (2013) 1898–1905.
[52] D. Keysers, F. Shafait, T. M. Breuel, Document image zone classification-a simple high-performance approach, in: Conference on Computer Vision Theory and Applications, 2007.
[53] N. Stamatopoulos, B. Gatos, T. Georgiou, Page frame detection for double page document images, in: International Workshop on Document Analysis Systems, 2010, pp. 401–408.
[54] T. Pavlidis, J. Zhou, Page segmentation and classification, CVGIP: Graph. Models Image Process. 54 (6) (1992) 484–496.
[55] M. Javed, P. Nagabhushan, B. Chaudhuri, Automatic extraction of correlation-entropy features for text document analysis directly in run-length compressed domain, in: International Conference on Document Analysis and Recognition, 2015, pp. 1–5.
[56] N. Otsu, A threshold selection method from gray-level histograms, Automatica 11 (285–296) (1975) 23–27.
[57] R.F. Moghaddam, M. Cheriet, A multi-scale framework for adaptive binarization of degraded document images, Pattern Recognit. 43 (6) (2010) 2186–2198.
[58] R.F. Moghaddam, M. Cheriet, AdOtsu: an adaptive and parameterless generalization of Otsu's method for document image binarization, Pattern Recognit. 45 (6) (2012) 2419–2431.
[59] B. Gatos, I. Pratikakis, S.J. Perantonis, Adaptive degraded document image binarization, Pattern Recognit. 39 (3) (2006) 317–327.
[60] M.A. Ramírez-Ortegón, L.L. Ramírez-Ramírez, V. Märgner, I.B. Messaoud, E. Cuevas, R. Rojas, An analysis of the transition proportion for binarization in handwritten historical documents, Pattern Recognit. 47 (8) (2014) 2635–2651.
[61] L.J. Latecki, R. Lakämper, Convexity rule for shape decomposition based on discrete contour evolution, Comput. Vis. Image Underst. 73 (3) (1999) 441–454.
[62] M.T. Parvez, S.A. Mahmoud, Arabic handwriting recognition using structural and

syntactic pattern attributes, Pattern Recognit. 46 (1) (2013) 141–154.
[63] X. Wang, B. Feng, X. Bai, W. Liu, L.J. Latecki, Bag of contour fragments for robust shape classification, Pattern Recognit. 47 (6) (2014) 2116–2125.
[64] X. Bai, C. Rao, X. Wang, Shape vocabulary: a robust and efficient shape representation for shape matching, IEEE Trans. Image Process. 23 (9) (2014) 3935–3949.
[65] D. K. Prasad, C. Quek, M. K. Leung, S.-Y. Cho, A parameter independent line fitting method, in: Asian Conference on Pattern Recognition, 2011, pp. 441–445.
[66] S. Belongie, J. Malik, J. Puzicha, Shape matching and object recognition using shape contexts, IEEE Trans. Pattern Anal. Mach. Intell. 24 (4) (2002) 509–522.
[67] L. Schomaker, L. Vuurpijl, Forensic Writer Identification: A Benchmark Data Set and A Comparison of Two Systems, Technical Report, NICI, Nijmegen, 2000.
[68] U.-V. Marti, H. Bunke, The IAM-database: an English sentence database for offline handwriting recognition, Int. J. Doc. Anal. Recognit. 5 (1) (2002) 39–46.
[69] H. Moon, P.J. Phillips, Computational and performance aspects of pca-based face-recognition algorithms, Perception 30 (3) (2001) 303–321.
[70] N. Arvanitopoulos, S. Süsstrunk, Seam carving for text line extraction on color and grayscale historical manuscripts, in: International Conference on Frontiers in Handwriting Recognition, 2014, pp. 726–731.
[71] A. Brink, M. Bulacu, L. Schomaker, How much handwritten text is needed for text-independent writer verification and identification, in: International Conference on Pattern Recognition, 2008, pp. 1–4.
[72] G. Ghiasi, R. Safabakhsh, Offline text-independent writer identification using codebook and efficient code extraction methods, Image Vis. Comput. 31 (5) (2013) 379–391.
[73] T. Van der Zant, L. Schomaker, K. Haak, Handwritten-word spotting using biologically inspired features, IEEE Trans. Pattern Anal. Mach. Intell. 30 (11) (2008) 1945–1957.
[74] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, A. Oliva, Learning deep features for scene recognition using places database, in: Advances in Neural Information Processing Systems, 2014, pp. 487–495.
[75] R.G. von Gioi, J. Jakubowicz, J.-M. Morel, G. Randall, LSD: a fast line segment detector with a false detection control, IEEE Trans. Pattern Anal. Mach. Intell. 4 (2008) 722–732.
[76] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: spatial pyramid matching for recognizing natural scene categories, in: Computer Vision and Pattern Recognition, vol. 2, 2006, pp. 2169–2178.
[77] L. Fei-Fei, P. Perona, A Bayesian hierarchical model for learning natural scene categories, in: Computer Vision and Pattern Recognition, vol. 2, 2005, pp. 524–531.

Sheng He received his B.S. and M.S. degrees both from Northwestern Polytechnical University, Xi'an, China, in 2009 and 2012, respectively. He is currently pursuing the Ph.D. degree in the Artificial Intelligence Department, University of Groningen, the Netherlands. His research interests include pattern recognition, image processing and handwritten document analysis.

Lambert Schomaker is a full professor in Artificial Intelligence at the University of Groningen and the director of its AI institute ALICE since 2001. His main interest is in pattern recognition and machine learning problems, with applications in handwriting recognition problems. He has contributed to over 160 peer-reviewed publications in journals and books (h=17/ISI, h=39/Google Citations). His work is cited in 23 patents. In recent years his focus is on continuous-learning systems and bootstrapping problems, where learning starts with very few examples. Prof. Schomaker is a senior member of IEEE, member of the IAPR and is a member of a number of Dutch research programme committees in e-Science (NWO), Computational Humanities (KNAW), Computational science and energy (Shell/NWO/FOM). He received IBM Faculty Awards (2011, 2012) for the Monk word retrieval system in historical manuscript collections using high-performance computing.