A Polar Stroke Descriptor for Classification of Historical Documents

Sheng He and Lambert Schomaker
Institute of Artificial Intelligence and Cognitive Engineering (ALICE), University of Groningen
Nijenborgh 9, Groningen, the Netherlands
Email: heshengxgd@gmail.com, L.Schomaker@ai.rug.nl

***Abstract*—This paper presents a method for extracting a local scale- and rotation-invariant descriptor called *Polar Stroke Descriptor* (PSD) from the strokes in handwritten documents. Our descriptor captures the stroke-length distribution on a reference point and results in a robust feature vector. Furthermore, we develop a compact stroke representation termed as *strokelets* inspired by the classical Bag of Words model. We apply the proposed representation to historical document dating and the experimental results show that the proposed method achieves state-of-the-art performance on the MPS database.**

I. Introduction

Strokes are the atomic elements in handwritten documents which contain glyph information and individual writing style. Extracting features of strokes is an important phase for character recognition [1], word-spotting [2], writer identification [3], [4] and handwritten document analysis [5]. The most popular feature for character recognition is the gradient direction histogram [1], for word spotting it is the SIFT [6] with the bag-of-word framework [7], and for writer identification it is the Hinge [8], [3], Quill [9], Δ^nHinge [4] and chain code [10] feature. Most of these statistical features represent the distribution of local stroke orientation/direction on the stroke boundaries. These features are simple and capture the statistical information in the entire document image. However, they fail to describe single stroke shapes, leading to a limited application area. In this paper, we aim to design a general local stroke descriptor for handwritten document analysis.

Two factors should be considered when designing a local stroke descriptor in handwritten images: Scale-invariance and Rotation-invariance. It has been shown that scale- and rotation-invariance are important for image matching or object recognition in natural scenes [6]. It is the same in handwritten documents, because in the real-world, documents are always digitized in different scales, and sometimes with a rotation angle. In addition, the descriptor should contain the ink-width or ink-length information. It was shown in [9] that ink-width can capture the pen properties.

In this paper, we propose a novel stroke descriptor named *Polar Stroke Descriptor (PSD)*. It is a local descriptor to describe the coarse distribution of the stroke length in every direction from 0 to 360 degrees with respect to a given reference point inside the stroke ink inspired by [9], [11]. The descriptor has three advantages: (i) It is a local descriptor for single stroke, and easy to be extended to a global feature using the bag-of-word framework. (ii) It is a scale-invariant and rotation-invariant descriptor. (iii) It contains the information of both stroke width and stroke length. Furthermore, we develop a new stroke representation called *Strokelets* inspired by [12]. The connected handwritten texts are decomposed into sub-strokes which are described by PSD on several sampling points. A codebook is trained using the Kohonen SOM 2D method and a statistical histogram of sub-strokes in a handwritten document is computed based on the learned codebook to represent the document.

II. Related work

The features used in handwritten document analysis have typically been categorized into two classes: statistical features and codebook-based features. Several statistical features have been proposed in the last two decades. The histogram of gradient direction is the widely used feature for handwritten digit recognition [1] and handwritten character recognition [13]. The SIFT feature, which has been successfully used in the computer vision field, was also applied for the word-spotting problem [7], [14]. In [15], the joint probability distribution of the angle combination of two "hinged" edge fragments was proposed for writer identification, which was termed as the "edge-Hinge" feature. This method has been extended to the contour-Hinge probability distribution [3] which computes a Hinge kernel on the contours of texts, Quill-Hinge [9] which combines the ink width with the contour-Hinge feature, and $\Delta^n Hinge$ [4] which is a rotation-invariant feature based on the contour-Hinge but incorporates the derivative between several points along the ink contours.

The codebook-based features are inspired by the Bag of Words framework [16] used in computer vision, which is useful in the case that some local elements are extracted from the images, but they can not be directly used to compute the similarity between these images. A codebook is learned from the local elements extracted from the entire data set in order to capture the general information. Finally, the feature vector of each image can be determined by computing the occurrence histogram of the members of the codewords. The commonly used codebook-based feature for handwritten images is the Bag of Words model powered by SIFT descriptors for word-spotting [7], [14]. In writer identification, several local elements have been proposed to represent the handwritten text. In order to capture features of the pen-tip trajectory which con-

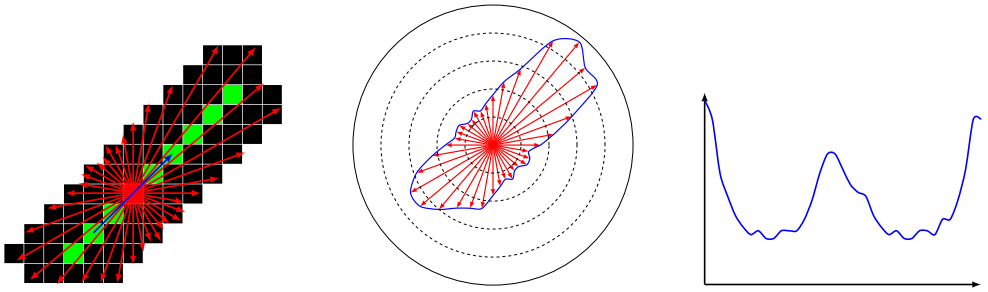


Fig. 1. An illustration of the polar stroke descriptor on a reference point (the red point in the left and middle figures). The red arrows are the stroke lengths computed in every direction in the discrete set $\mathcal{D}$. The blue arrow in the left figure shows the dominant direction φ_j which is the tangential direction on the skeleton line (green line). The middle figure shows the length distribution as a polar distribution (φ,r), and the right figure shows the normalized distribution of radius lengths starting from φ_j. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

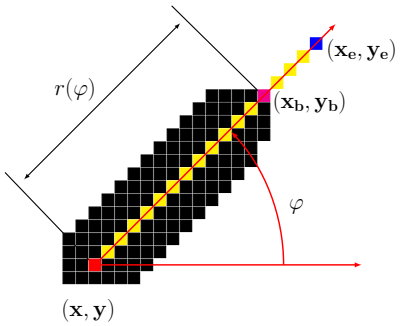


Fig. 2. Illustration of the stroke length computation on the direction φ. The red pixel is the reference point, and the blue one is the end pixel with direction φ. The pink pixel is the first background pixel that is hit when following a Bresenham path from the starting pixel (the red one) to the end pixel (the blue one). The length of the stroke on the direction φ is the Euclidean distance from the red pixel to the pink pixel.

tains valuable writer-specific information, Schomaker et al. [8] considered the COnnected-COmponent COntours (CO^3) as the basic elements. This approach was extended to fragmented CO^3s [3], [17] which are ink-blob shapes generated by the writers. Bulacu et al. [18] showed that a pixel-based (patch) codebook yields similar results in writer identification.

III. Our Approach

In this section, we first describe how to build the polar stroke descriptor given a reference point and then we extend it to the strokelets representation.

A. The Polar Stroke Descriptor (PSD)

As a key contribution, we propose a novel descriptor for handwritten strokes, the *Polar Stroke Descriptor*, that expresses the configuration of the entire stroke relative to the reference point. Consider a point p_i on the skeleton line inside the stroke ink, we firstly compute the stroke length in every direction φ in a discrete set $\mathcal{D}$ which is defined as: $\mathcal{D} = \left\{2\pi k/N; k \in \{0, \cdots, N-1\}\right\}$. N is the number of directions we considered and we empirically set it to 120.

The stroke length from the reference point to the ink boundary in the direction φ is computed by the approach in [9], which is based on Bresenham's algorithm [19]. The Bresenham's method constructs an approximated linear path of pixels between the given starting and end pixels in an image. In this paper, the starting pixel is the reference point $p_i = (x, y)$ and the end point (x_e, y_e) is found in the direction φ by:

$$\begin{aligned} x_e &= x + l * \cos(\varphi) \\ y_e &= y + l * \sin(\varphi) \end{aligned} \tag{1}$$

Here, the parameter l signifies the maximum measurable length. The length of the stroke can be measured by the trace length on the Bresenham path starting from the reference point $p_i = (x, y)$ towards the end point (x_e, y_e). The trace stops if a background (white) pixel (x_b, y_b) is hit and the trace length $r(\varphi)$ is then computed as the distance from the reference point p_i to this background pixel (x_b, y_b) by the Euclidean measure:

$$r(\varphi) = \sqrt{(x - x_b)^2 + (y - y_b)^2} \tag{2}$$

Fig. 2 gives an illustration of this method.

The polar stroke descriptor of the reference point p_i is defined as the normalized distribution of the stroke length computed by Eq. 2 in every direction in the discrete set $\mathcal{D}$:

$$PSD(p_i) = \{f(\varphi_j), \cdots, f(\varphi_{N-1}), f(\varphi_0), \cdots, f(\varphi_{j-1})\} \tag{3}$$

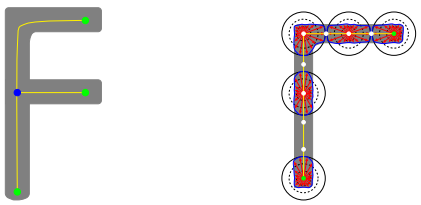


Fig. 3. A toy example of stroke extraction in figure (a) and stroke feature computation in figure (b). The yellow lines are the skeleton lines of the strokes. In figure (a), the green points are the end points on the skeleton line and the blue one is the fork point. In this case, the character can be segmented into three sub-strokes according to the fork and end points. In figure (b), the white points are the 8 sampling points on a sub-stroke between two end points (green points). All the sampling points including the end points are used as the reference points of PSD. Note that each solid circle in Figure (b) is plotted for showing the computation of the PSD on the corresponding position, and the radius is l. We only display five PSDs for better visualization.

Fig. 4. Images of charters in the MPS database [24].

Here $f(\varphi) = r(\varphi)/\sum_{n=0}^{N-1} r(\varphi_n)$ is the normalized length in the direction φ and φ_j is the dominant direction of p_i in order to achieve the invariance to stroke rotations. The tangential direction of the point p_i can be considered as the dominant direction which gives the most stable results (see the blue arrow in the left figure in Fig. 1). We estimate the dominant direction φ_j on the skeleton line using the method proposed in [9]. Fig. 1 gives an illustration of the length distribution on a reference point and the corresponding feature.

The addtional advantage of the proposed feature detection is that it can be used for junction detection if the reference points lie on the structure points of the strokes. The PSD in this case is specialized to junction feature [20].

B. Strokelets

Strokelets were first proposed in [12] for text recognition in natural scenes, which are the representation consisting of a set of multi-scale mid-level primitives, each of which under ideal conditions represents a stroke shape. In this paper, we apply this concept in handwritten document images and use the PSD to describe each stroke. Stroke extraction is an important research problem in handwritten documents and much research has been devoted to it [21], [22]. Here, we use a simple method to partition the connected texts into meaningful sub-strokes. Based on the assumption that the basic geometric elements that possess the desirable invariant properties are the fork points in 2D shapes, the skeleton produced by a thinning algorithm is cut into line segments from fork points and end points on the binarized handwritten images and each skeleton line segment corresponds to a sub-stroke. A toy example is shown in Fig. 3(a), in which the character is segmented into three sub-strokes.

For each extracted sub-stroke, we describe it using the proposed PSD as follows: we sample 10 reference points (including the two end points) equidistantly on the skeleton line (see Fig. 3(b)) and then compute 10 PSDs based on the reference points individually. The sub-stroke feature is a concatenation of the 10 PSDs on each sampling point.

In order to build a global feature for a handwritten document, the extracted sub-stroke features should be encoded into a new space spanned by a stroke codebook. In this paper, we use the Kohonen SOM 2D method, which has been successfully used for codebook generation in off-line text-independent writer identification [3], [23], to learn the strokelets codebook. The width of the Kohonen SOM codebook is set to 25 and finally the size of the codebook is $25 \times 25 = 625$.

IV. Experiments

Although our proposed feature can be used for many applications of handwritten document analysis, in this paper we apply the strokelets representation for historical document dating [24], which is a challenging problem compared to writer identification.

A. Dataset

The MPS dataset [24] contains historical documents from four cities representing the four corners of the Medieval Dutch language area: Arnhem, Leuven, Leiden and Groningen. Each document is dated and was written within five years before or after one of the quarter century years, from the period under consideration here (1300-1550), such as 1300, 1325, ..., 1550. We designate each of these quarter century years as a 'key year'. There are 1706 documents divided into 11 key years. Fig. 4 shows several images of charters in the MPS database. The MPS dataset is designed to estimate the year of an undated manuscript based on the general trend of writing styles in a certain period (± 5 around each key year). In this paper, we use the proposed feature to describe the writing style in each document.

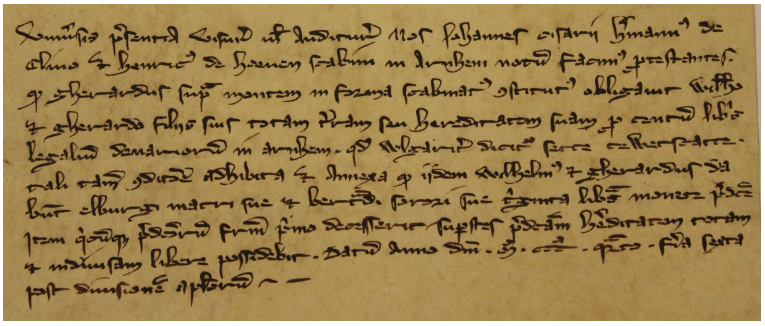
B. Visual results

In this section, we provide a visual result about our proposed feature. Given a historical document, we extract features of all the sub-strokes and use K-means to generate 50 clusters with the χ^2 distance function. Fig. 5(b) shows randomly selected 10 instances of 10 clusters in the given historical document (Fig. 5(a)).

We can see from Fig. 5 that our proposed feature is efficient to capture the structural information of the sub-strokes. Compared to Fraglets [3] which is used for writer identification, our stroke feature does not need any segmentation or line detection which are challenging problems in historical document images.

C. Historical document dating

Historical document dating can also be considered as a classification problem, in which documents from each key



(a) A historical document from the MPS data set.

(b) Instances in several randomly selected clusters.

Fig. 5. In figure (b), each row shows 10 instances of one cluster generated by K-means in the historical document of figure (a). The red lines in each sub-stroke patch are the skeleton lines of the stroke ink, and white points are the sampling points.

TABLE I
THE DATE (YEAR) CLASSIFICATION ACCURACY(%) OF THE k-NN CLASSIFIER ON THE MPS DATABASE.

Method	k=10	k=20	k=50	k=100
Δ^1Hinge [4]	53.3	50.1	45.6	41.2
Quill [9]	54.6	51.4	44.3	37.4
QuillHinge [9]	59.9	57.3	49.9	42.3
Hinge [3]	60.4	57.1	49.8	42.9
Strokelets	**60.7**	**58.7**	**55.3**	**48.2**

year belong to the same class. Classification was performed by a k nearest neighbors classifier (k-NN), which was also used in scene classification [25]. Given an undated historical document, the k-NN first looks for the k nearest neighbors of the undated document within the labeled database. Then, it assigns to the undated query document the year of the class (key year) the most represented within the k nearest neighbors. Performances of classification using the k-NN with different k are presented in Table I. As shown in this table, the proposed method performs better than all of state-of-the-art approaches. The possible reason is that the Hinge [3], Quill [9], QuillHinge [9] and Δ^1Hinge [4] are scale-sensitive features.

Besides the traditional classification accuracy, we also use the Mean Absolute Error (MAE) and the Cumulative Score (CS) measurements following [24] to evaluate the performance of historical document dating using the proposed feature. The MAE is typically defined as:

$$MAE = \sum_{j=1}^{N} |\overline{K(y_i)} - K(y_i)|/N \tag{4}$$

where $\overline{K(y_i)}$ is the estimated year of the input document y_i, $K(y_i)$ is the corresponding ground truth, and N is the number of test documents. The CS is defined as:

$$CS(\alpha) = N_{e \leq \alpha}/N \times 100\% \tag{5}$$

where $N_{e \leq \alpha}$ is the number of test documents on which the year estimation makes an absolute error no higher than α years.

Table II presents the results in term of the MAE and CS with error level $\alpha = 50$ using the k-NN classifier with different k. According to this table, the proposed method gets much better results than all of state-of-the-art approaches, and the best result is obtained when $k = 20$. The CS scores with $k = 50$ are shown in Fig. 6, which shows that scores of the proposed strokelets are higher than others on different error levels.

V. CONCLUSION

In this paper, we have proposed a new scale- and rotation-invariant local stroke descriptor termed as *Polar Stroke De-*

TABLE II
THE MAE AND CS OF THE k-NN CLASSIFIER.

Method	k=10		k=20		k=50		k=100	
	MAE	CS(α=50)	MAE	CS(α=50)	MAE	CS(α=50)	MAE	CS(α=50)
Quill [9]	28.1	83.5%	30.1	82.3%	34.9	78.8%	39.1	75.2%
Δ^1Hinge [4]	28.6	84.2%	29.9	83.1%	31.7	81.8%	35.6	78.1%
QuillHinge [9]	22.5	87.8%	23.6	87.5%	28.0	85.0%	33.3	80.5%
Hinge [3]	22.6	86.7%	23.9	85.9%	27.6	83.9%	31.9	80.5%
Strokelets	**22.1**	**87.8%**	**20.9**	**88.5%**	**22.4**	**87.9%**	**26.7**	**85.5%**

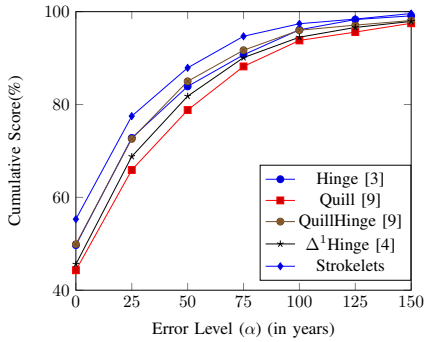


Fig. 6. Cumulative probability of $p(MAE \leq \alpha)$ on the error level (α) from 0 to 150 years of the k-NN classifier with $k = 50$.

scriptor and the strokelets representation. The PSD is particularly useful due to its distinctiveness which is achieved by assembling a high-dimensional feature vector representing the stroke-length in each direction.

This paper has also presented a method for using strokelets for historical document dating. The method we have described uses the k nearest neighbor classifier, and the performance of the proposed method is better than state-of-the-art approaches. Other potential applications include writer identification, word-spotting and handwritten character recognition.

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 Petros Samara and Jan Burgers for their contributions of constructing the MPS data set.

REFERENCES

[1] C.-L. Liu, K. Nakashima, H. Sako, and H. Fujisawa, "Handwritten digit recognition: investigation of normalization and feature extraction techniques," *Pattern Recognition*, vol. 37, no. 2, pp. 265–279, 2004.
[2] J.-P. Van Oosten and L. Schomaker, "Separability versus prototypicality in handwritten word-image retrieval," *Pattern Recognition*, vol. 47, no. 3, pp. 1031–1038, 2014.
[3] M. Bulacu and L. Schomaker, "Text-independent writer identification and verification using textural and allographic features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 4, pp. 701–717, 2007.
[4] S. He and L. Schomaker, "Delta-n hinge: rotation-invariant features for writer identification," in *ICPR*, 2014, pp. 2023–2028.
[5] K. Chen, H. Wei, J. Hennebert, R. Ingold, and M. Liwicki, "Page segmentation for historical handwritten document images using color and texture features," in *ICFHR*, 2014, pp. 488–493.
[6] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[7] M. Rusiñol, D. Aldavert, R. Toledo, and J. Lladós, "Efficient segmentation-free keyword spotting in historical document collections," *Pattern Recognition*, vol. 48, no. 2, pp. 545–555, 2015.
[8] L. Schomaker and M. Bulacu, "Automatic writer identification using connected-component contours and edge-based features of uppercase Western script," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 4, pp. 787–798, 2004.
[9] A. Brink, J. Smit, M. Bulacu, and L. Schomaker, "Writer identification using directional ink-trace width measurements," *Pattern Recognition*, vol. 45, no. 1, pp. 162–171, 2012.
[10] I. Siddiqi and N. Vincent, "Text independent writer recognition using redundant writing patterns with contour-based orientation and curvature features," *Pattern Recognition*, vol. 43, no. 11, pp. 3853–3865, 2010.
[11] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 4, pp. 509–522, 2002.
[12] C. Yao, X. Bai, B. Shi, and W. Liu, "Strokelets: A learned multi-scale representation for scene text recognition," in *CVPR*, 2014, pp. 4042–4049.
[13] C.-L. Liu and C. Y. Suen, "A new benchmark on the recognition of handwritten Bangla and Farsi numeral characters," *Pattern Recognition*, vol. 42, no. 12, pp. 3287–3295, 2009.
[14] J. Almazán, A. Gordo, A. Fornés, and E. Valveny, "Word spotting and recognition with embedded attributes," in *IEEE Trans. Pattern Anal. Mach. Intell.*, 2014.
[15] M. Bulacu and L. Schomaker, "Writer style from oriented edge fragments," in *Computer Analysis of Images and Patterns*. Springer, 2003, pp. 460–469.
[16] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *CVPR*, 2005, pp. 524–531.
[17] M. Bulacu, L. Schomaker, and A. Brink, "Text-independent writer identification and verification on offline Arabic handwriting," in *ICDAR*, 2007, pp. 769–773.
[18] M. Bulacu and L. Schomaker, "A comparison of clustering methods for writer identification and verification," in *ICDAR*, 2005, pp. 1275–1279.
[19] D. Hearn and M. P. Baker, *Computer graphics, C version*. Prentice Hall Upper Saddle River, 1997, vol. 2.
[20] S. He, M. Wiering, and L. Schomaker, "Junction detection in handwritten documents and its application to writer identification," *Pattern Recognition*, doi:10.1016/j.patcog.2015.05.022.
[21] V. Pervouchine, G. Leedham, and K. Melikhov, "Three-stage handwriting stroke extraction method with hidden loop recovery," in *ICDAR*, 2005, pp. 307–311.
[22] X. Liu, Y. Jia, M. Tan *et al.*, "Geometrical-statistical modeling of character structures for natural stroke extraction and matching," in *IWFHR*, 2006.
[23] L. Schomaker, K. Franke, and M. Bulacu, "Using codebooks of fragmented connected-component contours in forensic and historic writer identification," *Pattern Recognition Letters*, vol. 28, no. 6, pp. 719–727, 2007.
[24] S. He, P. Samara, J. Burgers, and L. Schomaker, "Towards style-based dating of historical documents," in *ICFHR*, 2014, pp. 265–270.
[25] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput, Vis.*, vol. 42, no. 3, pp. 145–175, 2001.