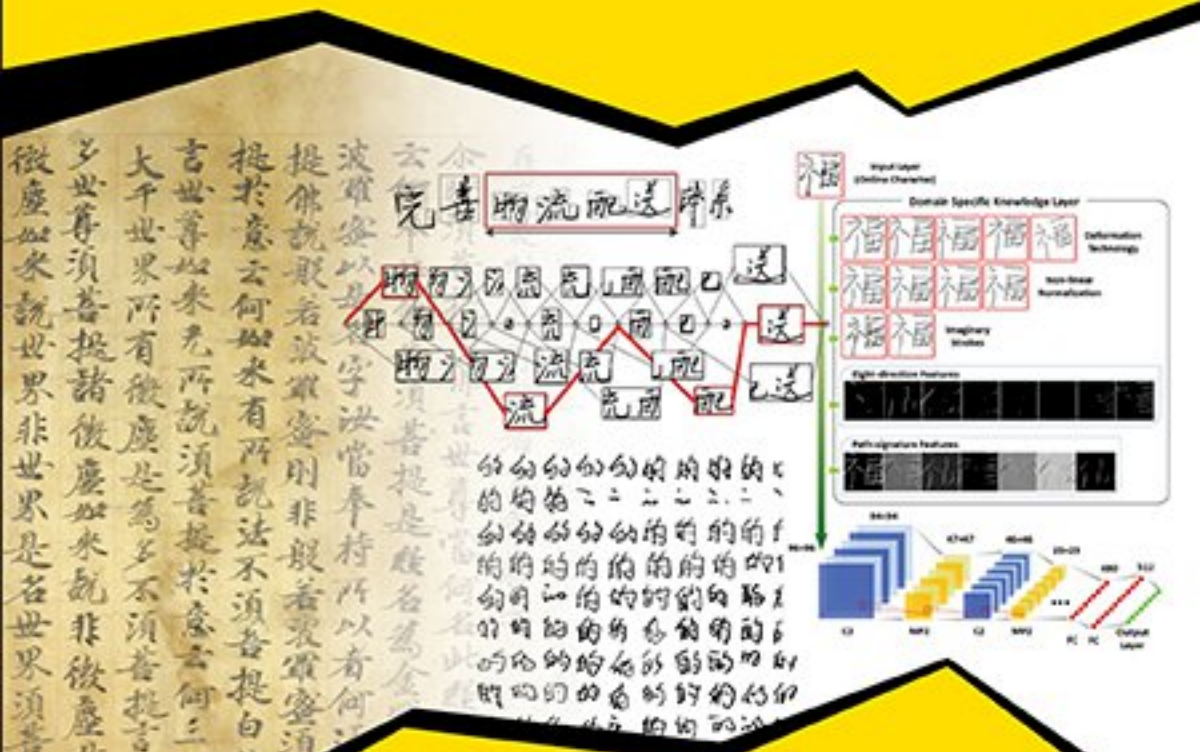
Series on Language Processing, Pattern Recognition, and Intelligent Systems — Vol. 2

Advances in Chinese Document and Text Processing

Edited by

Cheng-Lin Liu • Yue Lu



World Scientific

Chapter 1

Off-line Text-independent Writer Identification for Chinese Handwriting

Yu-Jie Xiong and Yue Lu

Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University, Shanghai 200241, China

ylu@cs.ecnu.edu.cn

Encouraged by the strong requirements of the information security, the rapid development of biometrics becomes a new focus in both academic and industrial research. Writer identification is a branch of behavioral biometrics using handwriting with a natural writing attitude as the individual characteristics for identification. Off-line text-independent writer identification is to identify a person based on the static handwritten data with unrestricted text content. We propose an effective method using the contour-directional feature (CDF) combined with the modified SIFT for Chinese writer identification. The investigation demonstrates that both features are capable of describing the characteristic of handwriting. In the stage of the modified SIFT extraction, a simple connected-component based segmentation algorithm is used to segment the handwriting image into character regions, and the modified SIFT descriptors are extracted from character regions. Then, a codebook is constructed by K-means clustering. With the codebook, the occurrence histogram of the modified SIFT for each handwriting image is calculated. Both features are concatenated together to represent the characteristic of handwriting. Experimental results show that the proposed method is able to improve the performance and superior to other methods in terms of identification accuracy. On the HIT-MW Chinese handwriting dataset involving 240 writers, the Top-1 accuracy is 96.3%, and the Top-10 accuracy is 99.2%.

1. Introduction

The requirements of personal authentication have placed biometrics at the center of the academic and industrial research, as it is becoming a key aspect of information security.[1] Biometrics refers to analyzing the

biological phenomena and observations by means of statistical techniques for individual recognition. It is performed by comparing the biometric template measured from an unknown person with templates linked to known persons with certainty. Depending on the adoptive traits, biometrics can be categorized into physiological biometrics and behavioral biometrics. Physiological biometrics identifies a person using a physical property of the human body (e.g. DNA, iris, fingerprint, face, and hand geometry), while behavioral biometrics considers individual traits of a person's behavior (e.g. voice, gait, signature, handwriting) for authentication. It is worth noting that handwriting is the most widespread carrier of personal behavioral information, and people have employed signatures as the legitimate means to verify an individual's identity for several centuries. Moreover, acquisition of handwriting is not invasive, and we always do some writing in our daily life which makes handwriting easy to get. For these reasons, handwriting with a natural writing attitude is an effective way to represent the individual characteristics, and plays an essential role in the set of biometric traits. In consideration of the variability of handwriting, writer identification is still an attractive but challenging research field.

1.1. *Writer identification vs. handwriting recognition*

Compared with writer identification, handwriting recognition is an older and broader research domain which has lasted for several decades.[2] The key of handwriting recognition is to obtain the invariant representation from a large number of manuscripts to eliminate the individual variations of handwriting.[3] However, the individual style of handwriting is the biometric trait for writer identification. Researchers try to find the specificity of writing style to achieve personal authentication. From this perspective, writer identification and handwriting recognition can be regarded as two different filters: handwriting recognition aims to get the text content from the low frequency domain (invariance) of handwriting, while writer identification tries to obtain the individual characteristic of the high frequency domain (variation). Although the targets of writer recognition and handwriting recognition are almost completely opposite, an interesting phenomenon indicates that handwriting recognition system can achieve better results with the writer adaptation.[4]

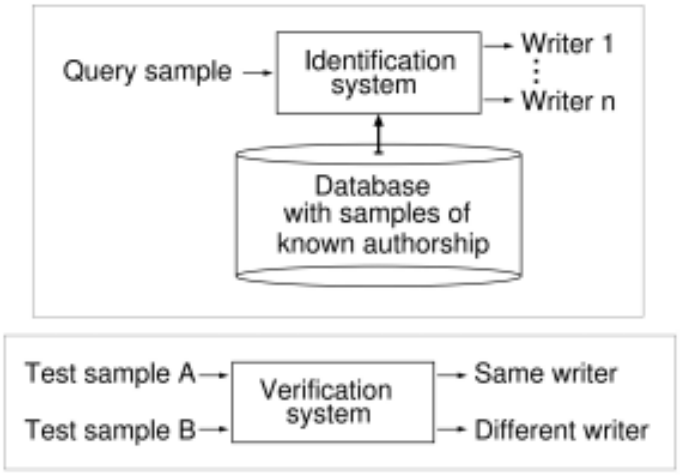


Fig. 1. A writer identification system and a writer verification system[3]

1.2. *Writer identification vs. writer verification*

Writer identification and writer verification are actually very similar. As shown in Fig. 1, writer verification involves a one-to-one comparison to determine whether two samples of handwriting are produced by the same person or not. Writer identification is to select the authentic author of handwriting from among a group of writers, and a sorted list of candidates is returned as the output. In addition, writer verification is also very similar with signature verification. The only difference between them is the text content of input. The former is a piece of handwriting, which contains several words or a few text lines. The text content of the reference is not related to the content of the query. The latter is a signature, which means the text content of the reference and the query is the same. Furthermore, writer verification assumes that handwriting is produced naturally,[5] whereas signature verification considers different conditions, including genuine and forged handwritings.[6]

1.3. *Text-dependent vs. text-independent*

On the basis of the content of handwriting, there are two different subcategories of writer identification: text-dependent and text-independent.[7] The text-dependent writer identification is similar to signature verification. It assumes that the references and the query share

the same text content. For text-dependent identification system, interactive tools are widely used to precisely find out the same characters or words. As a consequence, manual intervention makes its objectivity in doubt and its applicability is also restricted. Text-independent writer identification eliminates the restriction of text content and involves less manual work. It treats writer identification as a statistical classification problem rather than a template matching problem, so text-independent writer identification requires sufficient handwritten text of each writer to extract robust statistical feature for pattern representation. Therefore, the minimal amount of handwritten text which can satisfy the statistics assumption is of crucial importance for the text-independent writer identification.

1.4. *Recent progress in Chinese writer identification*

Writer identification is relevant to disciplines ranging from neuroscience to computer science, and it is an attractive but challenging research field which attracts a lot of interests from both academia and forensics. Plamondon and Lorrtte[7] summarized the progress of writer identification and signature verification in 1989. Continuing attentions and efforts have been devoted to reaching text-independent writer identification. Here we present several popular off-line text-independent approaches proposed since 2000, as a survey of recent developments for the topic of Chinese writer identification.

Zhu et al.[8] used the two-dimensional Gabor filtering technique to extract texture features and a weighted Euclidean distance classifier to fulfil the identification task. Shen et al.[9] improved the Gabor filters with the wavelet technique to reduce the excessive calculational cost and K-nearest neighbor (KNN) classifier was utilized to identify the writer. He and Tang[10] used both autocorrelation function and Gabor filters to extract the features, and weighted Euclidean distance classifier is used to match the extracted features. He et al.[11] presented hidden Markov tree model (HMTM) in wavelet domain for Chinese writer identification. Compared with the two-dimensional Gabor model, the HMTM not only achieved better identification performance but also greatly reduced the elapsed time. After that, they also presented a wavelet based method with generalized Gaussian density model.[12] Zhang et al.[13] proposed a hybrid method combining Gabor model with mesh fractal dimension. Li and Ding[14] proposed a histogram-based feature called as the grid microstructure feature (GMF) which was extracted from the edge image, and the similarity of different handwritings was measured with the improved

weighted Chi-squared metric. It is noted that they were the winners of the ICDAR 2011 writer identification contest.[15] However, the grid microstructure feature is sensitive to pen-width variation in practical situation. Xu et al.[16] proposed an inner and inter class variances weighted feature matching method to solve this problem. Wen et al.[17] characterized the frequent structures distribution of edge fragments on multiple scales to describe the writing style of Chinese handwriting, and applied Chi-squared distance as similarity measurement. Hu et al.[18] employed the SIFT descriptor to describe the local directional information of Chinese characters and the KNN classifier was used to identify the author of handwriting. Instead of hard voting, they also presented two coding strategies as improved fisher kernels and locality-constrained linear for feature coding.

2. The proposed method

We present a new writer identification method to identify the authentic writer of the query handwriting document using the contour-directional feature and the modified SIFT occurrence histogram. The flowchart of the proposed method is given in Fig. 2. The contour-directional feature is extracted from the contour image obtained by contour detection. The extraction of the modified SIFT occurrence histogram is divided into two stages: codebook generation and histogram calculation. First, a segmentation algorithm based on connected-component is used to segment the handwriting image into character regions. Then, the modified SIFT descriptors are extracted from the character regions. These obtained descriptors are used to create the codebook using K-means clustering. After that, the occurrence histogram of the modified SIFT descriptors for each handwriting is calculated with the codebook. The contour-directional feature and the modified SIFT occurrence histograms are concatenated together for similarity measurement. In the proposed method, weighted Chi-squared distance is utilized to calculate the distance of the contour-directional features and Manhattan distance is used to calculate the distance of the modified SIFT occurrence histograms. The two distances are accumulated to measure the final similarity.

2.1. *Contour-directional Feature*

It is important to extract appropriate features to represent handwriting image for similarity measurement. Bulacu et al.[19] proposed a feature of

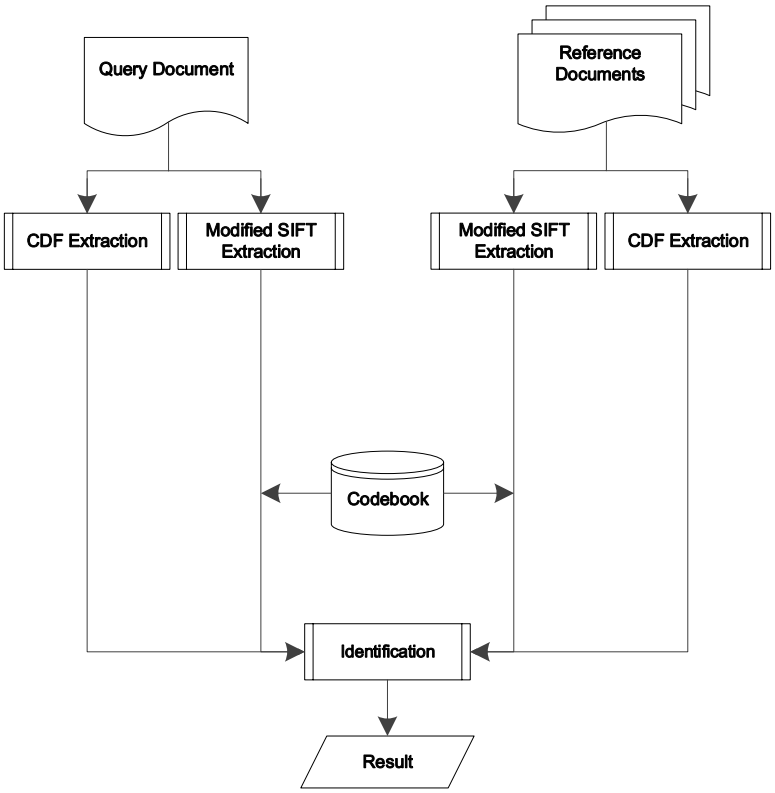
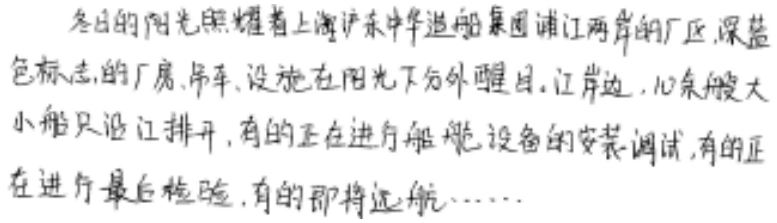
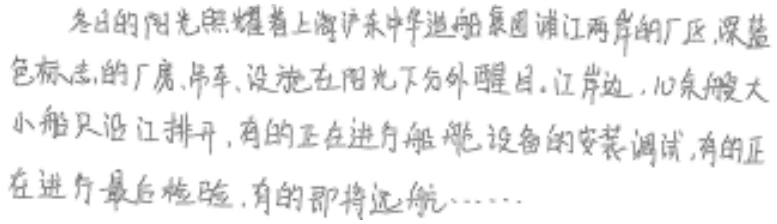


Fig. 2. System flow chart

the edge-hinge distribution which represents the changes in the direction of handwriting strokes. Though experimental results show that identification performance of edge-hinge feature outperforms all non-angular features, the edge-hinge feature only computes the joint probability distribution of the directions of the two edge in the neighborhood. In addition, Li and Ding[14] expanded the edge-hinge feature into grid microstructure feature for Chinese writer identification. Chinese is a hieroglyphic writing. Compared with alphabetic writing (like English), Chinese character has more complex stroke crossings, and single character is always located in a relative separate block region. Grid microstructure feature generated from a grid of variable size and records the positions of specific edge pixel pairs. It calculates the probability distribution of the microstructure by a series of moving grid windows. But according to the definition of the GMF, if the edge pixel



(a)



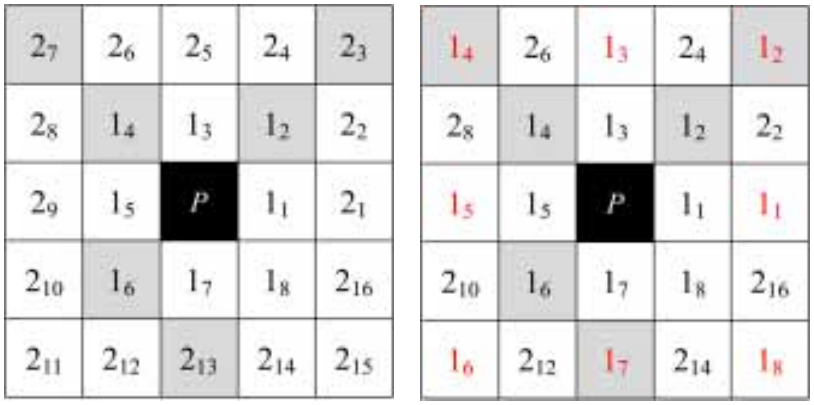
(b)

Fig. 3. Original Sample(a) and handwriting image after contour detection(b).

pairs with the same directions are located in the grids with different scales, they are regarded as different microstructures. From the perspective of directions, these pixel pairs are very similar to each others though them have different scales. Hence, we propose the contour-directional feature, which characterizes the writing style of handwriting by the distribution of pixel pairs based on the directional information. The contour-directional feature retains local information of the character, not only the relationships of stroke structures but also directions of pixel pairs.

The first step of contour-directional feature extraction is contour detection. We use the Sobel operators to extract the contour from the original handwriting image. An example of contour detection is shown in Fig. 3.

Then, we extract the specific pairs of edge pixel from the contour image. To obtain these pairs of edge pixel, the contour image is divided into a number of blocks of size $n \times n$, and the center of each block is a edge pixel. In each block, we find all the edge pixel pairs (α, β) which satisfy the



(a) Find specific edge pixel pairs (b) Redefine the index of pixels

Fig. 4. An example of the extraction of the CDF

following conditions:

$$\begin{cases} \alpha \text{ and } \beta \text{ are edge pixels,} \\ G(\alpha) = A_i,\ G(\beta) = B_j, \text{ and } A = B,\ i < j, \\ \text{If } G(\gamma) = A_k, \text{ and } i < k < j, \\ \text{then } \gamma \text{ is not the edge pixel.} \end{cases}$$

As shown in Fig. 4(a), it denoted a block of 5×5. The black square is the edge pixel P, and the gray squares are edge pixels connected to P. The pixel A around P is marked with the index $G(A) = Dis_i$, where Dis denotes the larger distance in the horizontal and vertical distance between A and P, and $1 \leq i \leq 8 * Dis$. This step is similar to the GMF.[14]

Then, we define the direction $Dir(A)$ of the pixel A in the block as:

$$Dir(A) = \arctan\left((A_y - P_y)/(A_x - P_x)\right)$$

Where (A_x, A_y) and (P_x, P_y) are the coordinates of A and P. Afterwards, we redefine the index of each pixel in the block according to the pixel's direction. The new index of A is denoted as $C(A)$. The naming rule of $C(A)$ is defined as:

$$\begin{cases} \text{If } Dir(A) \text{ is unique, then } C(A) = G(A) = dis_i; \\ \text{If } Dir(A) = Dir(A_1) = ... = Dir(A_n), \\ dis(An) < ... < dis(A1) < dis(A) \\ \text{then } C(A) = C(A_1) = ... = C(A_n) = G(A_n) = dis(A_n)_i; \end{cases}$$

As shown in Fig. 4(b), the changed indexes are labeled with red color. As a comparison, $(1_2, 1_4)$, $(1_4, 1_6)$, $(2_3, 2_7)$, $(2_7, 2_{13})$ in Fig. 4(a) are recorded

as the GMF, while the edge pixel pairs $(1_2, 1_4)$, $(1_4, 1_6)$, $(1_2, 1_4)$, $(1_4, 1_7)$ in Fig. 4(b) are recorded as the CDF.

Every edge pixel is surrounded by the block of size $n * n$, so we record the occurrence numbers of all the specific edge pixel pairs (α, β) of each block, and accumulate them to calculate the frequency histogram of pixel pairs after normalization. In this way, we acquire the contour-directional feature vector.

2.2. *Modified SIFT*

Lowe proposed Scale Invariant Feature Transform (SIFT)[20] in 2004, which has been successfully applied to object detection. The four major stages of SIFT are: (1) detection of scale-space extrema, (2) accurate keypoint localization, (3) orientation assignment, and (4) keypoint descriptor extraction. For the problem of writer identification, we hope that the features can describe the local patterns of stroke structures and use them to represent the characteristics of handwriting. However, SIFT has two shortcomings for the application of writer identification. First, the previous work was focused on extracting the SIFT descriptors from the whole handwriting image directly. It means that a part of descriptors obtained by the global extraction are located in the background area and are redundant. Second, SIFT is scale invariant, keypoints are detected from the different scale space. These characteristics are beneficial for object detection when the object in two image has different sizes. But considering the characteristics of handwriting image, the size of the character has no obvious change in the handwriting images of the same writer. In order to obtain SIFT descriptors appropriately, we propose some modifications for the original SIFT. The modified SIFT contains six major stages: (1) character region segmentation, (2) detection of scale-space extrema, (3) keypoint localization, (4) keypoint selection, (5) orientation assignment, and (6) modified descriptor extraction.

2.2.1. *Character region segmentation*

The modified SIFT extracts the features from character regions rather than the whole handwriting image, so we need to segment the handwriting image into character regions. Given a handwriting image I, the segmentation process is described as three steps:

a. I is converted to the binary image I_b using Otsu algorithm.

b. Connected-components in I_b are labeled and their average height H_a and average height W_a are calculated.

c. Connected-components C_i and C_j are merged if they meet the following conditions: (1) overlapping area of C_i and C_j is larger than 20% of the total area of C_i and C_j. (2) overlapping area of C_i and C_j is larger than 60% of the smaller area of C_i and C_j. (3) the vertical distance of centres of C_i and C_j is less than 25% of H_a. (4) the horizontal distance of centres of C_i and C_j is less than 25% of W_a.

After the segmentation, a handwriting image is divided into many character regions.

2.2.2. *Keypoint selection*

Compared with natural scene image, handwriting image lacks gray scale variation, so the original SIFT does not work well in the handwriting image. Based on the characteristics of handwriting, we add the step of keypoint selection to overcome this shortcoming. In general, the allographic information of character exists in the stroke of the character rather than the background area. Therefore, we apply the background point elimination to remove these useless keypoints. This procedure is performed according to the following criterions: If $S_n < n-1$, p is regarded as a keypoint in the background area; if $S_n \geq n-1$, p is regarded as a keypoint in the stroke area, where p is a detected keypoint, and S_n is the number of the black pixel in the $n \times n$ spatial neighbor grid of p. We remove the keypoints in the background area. Through the above operations, the remaining keypoints are credible representation of the individuality of handwriting. Fig. 5 is an example of the background point elimination. Fig. 5(a) is an original handwriting image, and Fig. 5(b) is the image after extrema detection, in which the colorized circles with different sizes are the potential keypoints in different scales. The red circles are keypoints in the background area, while the cyan ones are keypoints in the stroke area. After background point

(a) (b) (c)

Fig. 5. Keypoint selection processing. (a) original sample, (b) potential keypoints after keypoint localization, and (c) keypoints after background point elimination.

elimination, only the keypoints in the stroke area are remained in Fig. 5(c).

2.2.3. *The modified SIFT descriptor*

After keypoint selection, the modified SIFT descriptor of each keypoint is computed. The modified SIFT descriptor contains two parts: the original descriptor and the orientation of the keypoint. It is well known that by assigning a consistent orientation to each keypoint based on local image properties, the original SIFT descriptor is represented relative to this orientation and therefore achieves invariance to image rotation. However, the invariance to image rotation is not necessary for the issue of writer identification, and the orientation of the keypoint is discriminative to represent the writing style of handwriting, so we combine it with the original descriptor to create the modified SIFT descriptor. The orientation of the keypoint $L(x, y)$ is calculated as:[20]

$$\theta(x, y) = \tan^{-1}\left(\left(L(x, y+1) - L(x, y-1)\right) / \left(L(x+1, y) - L(x-1, y)\right)\right)$$

The range of $\theta(x, y)$ is $(-\pi, \pi)$, and it is quantized to 8 intervals. An example of the modified SIFT descriptor is shown as Fig. 6.

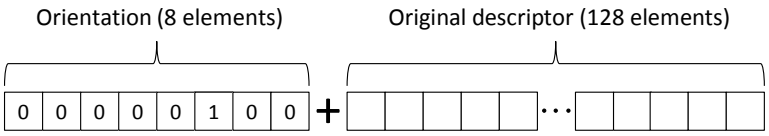


Fig. 6. An example of the modified SIFT descriptor

2.2.4. *Codebook generation and the modified SIFT occurrence histogram calculation*

Fig. 7 sketches the main modules of the codebook generation: After the step of the modified SIFT extraction, thousands of modified SIFT descriptors are extracted from handwriting. It is hard to calculate the similarity of different handwritings using those descriptors directly. To solve this problem, we utilize K-means clustering to cluster the descriptors into N classes as the codebook and represent each class by its center $(C_1, C_2 ... C_N)$. In order to keep the independence of the codebook, we use the training images to create the codebook. And N is set equal to 300 empirically in our experiments.

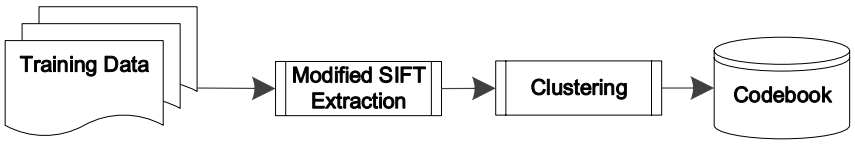


Fig. 7. Codebook generation

After the codebook generation, the modified SIFT occurrence histogram of each image can be calculated. For each SIFT descriptor, we calculate its nearest cluster center $C_i(1 < i \leq N)$ of the codebook based on Euclidean distance, and the occurrence counter corresponding to C_i is incremented by one. After all SIFT descriptors are calculated, the normalized occurrence histogram is treated as feature representation of handwriting.

2.3. *Identification*

Query image and the reference image are denoted as Q and R, let $CDF_Q = \{a_1, a_2, ..., a_N\}$ and $CDF_R = \{b_1, b_2, ..., b_N\}$ denote their contour-directional features, and let $SOH_Q = \{x_1, x_2, ..., x_M\}$ and $SOH_R = \{y_1, y_2, ..., y_M\}$ denote their modified SIFT occurrence histograms. Many existed methods based on minimum distance can be used for distance measure. In our method, weighted Chi-squared distance is used to calculate the distance D_C between CDF_Q and CDF_R:

$$D_C = \sum_{i=1}^{N} \frac{(a_i - b_i)^2}{(a_i + b_i)}$$

And Manhattan distance is used to calculate the distance D_S between SOH_Q and SOH_R:

$$D_S = \sum_{i=1}^{M} (|x_i - y_i|)$$

We summarize the two distances together to measure the final dissimilarity between Q and R:

$$D = D_C + D_S$$

For a given query handwriting, we calculate its distance to all reference handwritings, then a distance-based candidate list is obtained by sorting the results from the most similar to the least similar handwriting.

Table 1. Overview of the Experimental Datasets

Dataset	No. of documents	No. of queries	No. of references
Set-A	480	240	240
Set-B	1,704	853	853
LPAIS	400	200	200

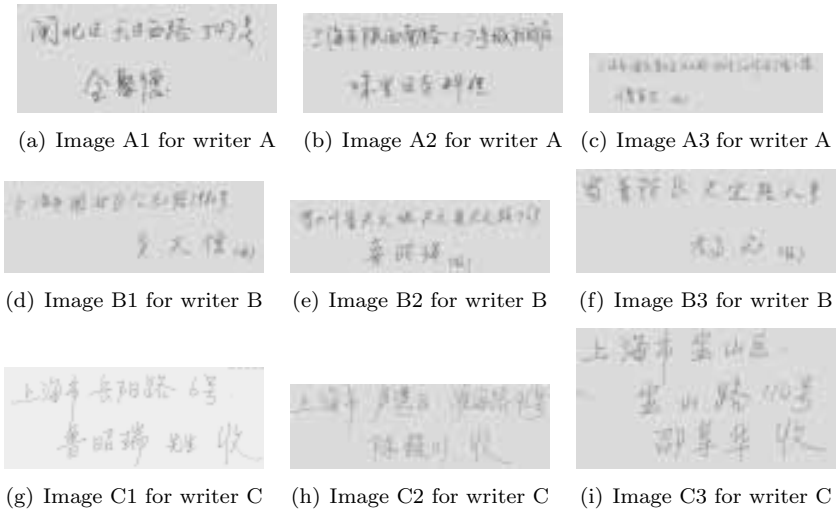
Note: 100 images of each reference set are used to generate the codebook.

3. Experimental Results

We evaluated the proposed method on the HIT-MW and our own LPAIS dataset. The Top-N criterion is used for evaluating the identification performance. For the TOP-N criterion, we consider a correct hit when at least one document image of the same writer is included in the N most similar document images.

HIT-MW[21] is built for the off-line Chinese handwritten text recognition, but it can also be used for the research of writer identification with the writer information list. HIT-MW includes 853 documents and each document consists of at least 200 characters. Two sub-datasets are generated from the HIT-MW dataset. Set-A contains 240 documents written by 240 writers. Every writer has one page in this dataset. Set-B contains all 853 documents of HIT-MW, but a few documents are created by the same writers. For simplicity, we assume that this dataset is created by 853 writers, and each writer only provides one page to simulate the situation of large number of writers. As shown in Tab. 1, Set-A contains 480 sub-document images, and Set-B contains 1706 images. In our experiments, each document image is segmented into two sub-images. One of them is used for reference, and the other one is used for query. A simple segmentation algorithm based on ground truth information is utilized to segment each image into two commensurate parts.

LPAIS is built for handwritten mail address analysis. It is collected from the practical mail images. LPAIS contains 400 handwritten mail address images collected from 20 Chinese writers and 20 images per writer. For each image, the address content is unrestricted. In most cases, the number of characters in the image is less than 30. Some mail address images are shown in Fig. 8. Compared with images in HIT-MW, we can see that the mail address images have various layouts and less characters. This implies that LPAIS is a more challenging dataset for the task of text-independent writer identification.



(a) Image A1 for writer A (b) Image A2 for writer A (c) Image A3 for writer A

(d) Image B1 for writer B (e) Image B2 for writer B (f) Image B3 for writer B

(g) Image C1 for writer C (h) Image C2 for writer C (i) Image C3 for writer C

Fig. 8. Some mail address images from three writers.

3.1. *Codebook size*

The size of the codebook has a significant impact on the performance of the modified SIFT. In this experiment, a range of codebook sizes are tested to find the optimal parameter of codebook size. For Set-A and Set-B, 100 samples of each reference set are used for codebook generation. Fig. 9 shows that the Top-1 accuracy of identification is improved with the increasing of codebook size. When the codebook size is larger than 300, the accuracy drops slightly. So the codebook size is set equal to 300 in the following experiments.

3.2. *CDF vs. GMF*

Experiments are performed to validate the performance of the CDF. For each query, the distances between it and all references are calculated. The writer of the reference corresponding to the minimal distance is the most likely writer of the query. Fig. 10 shows that the Top-1 accuracy and the Top-5 accuracy of the CDF are very close to those of the GMF on Set-A. When the number of writers becomes larger, the CDF performs better than the GMF on Set-B. It indicates that the CDF can be considered as an modification of the GMF, and its performance is more reliable.

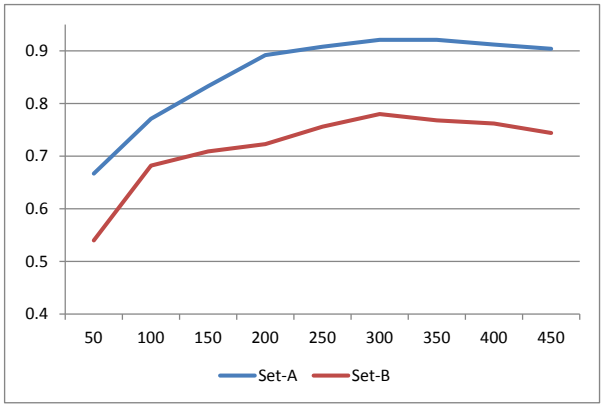


Fig. 9. The identification (Top-1) performance with different codebook sizes

Fig. 10. The identification performance comparison of the CDF and the GMF

3.3. *Modified SIFT vs. SIFT*

This experiment is used to compare the modified SIFT and the original SIFT. As shown in Fig. 11, we observe that the performance of the modified SIFT method drops a lot in Set-B. The Top-1 accuracy drops from 92.1%

to 78.0% and the Top-5 accuracy drops from 95.4% to 87.6%. These demonstrate that the modified SIFT is not very robust to the amount of writers. However, the modified SIFT is much better than original SIFT on both datasets. It shows that the modified SIFT profits from the additional orientation information. On the other hand, the segmentation of character regions also improves the identification accuracy by reducing the redundant keypoints.

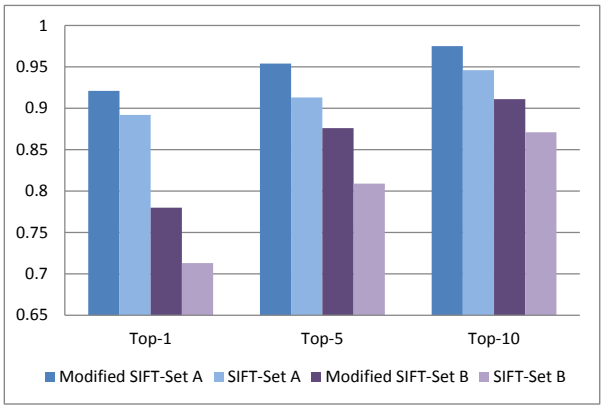


Fig. 11. The identification performance comparison of the modified SIFT and SIFT

3.4. *Combination of the CDF and the modified SIFT*

Fig. 12 and Fig. 13 show the performance of different features in two datasets. As shown in the both figures, compared with the sole CDF and the modified SIFT, the combination of both features dramatically improves the performance. It demonstrates that the proposed method can work well on the Chinese handwriting. Although the performance of the modified SIFT is not as good as that of the CDF, the fusion of them can improve the performance. The Top-1 accuracy grows from 95.4% to 96.2% and the Top-5 accuracy grows from 97.9% to 98.8% on Set-A. A possible reason is that the CDF and the modified SIFT characterize handwriting from different aspects, so the combination of the two features contributes to enhancing the performance effectively.

The results of some state-of-the-art methods on the same datasets are also provided for comparison. In Tab. 2, we compare the proposed method with four existing methods, the grid microstructure feature[14] , the edge-hinge distribution,[19] the multi-scale edge-hinge combinations[22] and

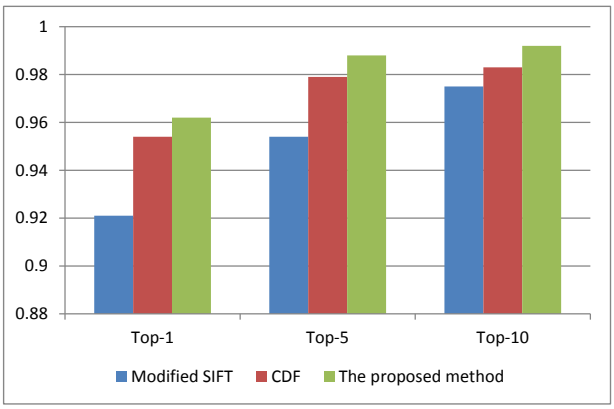


Fig. 12. The identification performance of different features on Set-A

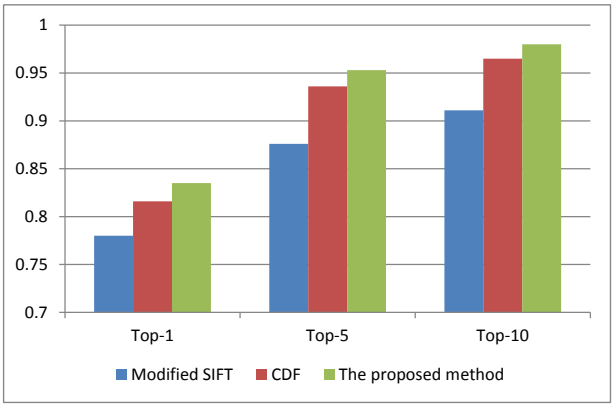


Fig. 13. The identification performance of different features on Set-B

Table 2. The Top-N performance of different methods on Set-A

Top-N / Methods	Top-1	Top-5	Top-10
Grid microstructure feature[14]	95.0%	98.3%	98.8%
Edge-hinge distribution[19]	91.7%	-	-
Edge-hinge combinations[22]	93.8%	-	-
Edge structure coding[17]	95.4%	-	-
The proposed method	**96.3%**	**98.8%**	**99.2%**

Table 3. The Top-N performance of different methods on Set-B

Top-N / Methods	Top-1	Top-5	Top-10
Grid microstructure feature[14]	80.3%	92.5%	95.7%
The proposed method	**83.5%**	**95.3%**	**98.0%**

the edge structure coding[17] on Set-A. For comparison, we also implement the grid microstructure feature[14] for on Set-B. The results in Tab. 3 show that the Top-1, Top-5 and Top-10 accuracy of the proposed method is better than that of the GMF.

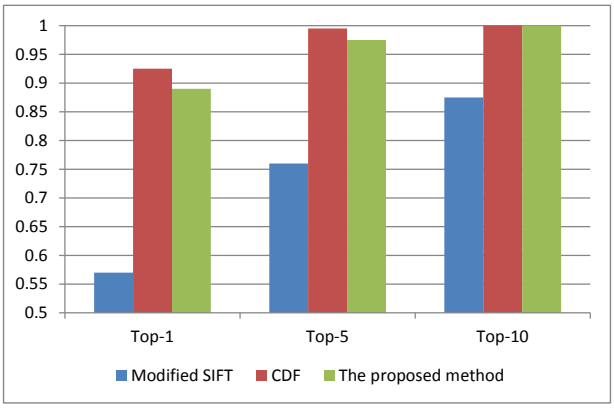


Fig. 14. The identification performance of different features on LPAIS

To further investigate the robustness, the proposed method is tested on the LPAIS dataset. It is noted that mail address images of LPAIS have less characters and quite different layouts. Fig. 14 shows the writer identification performance of different features on this dataset.

The Top-1 accuracy of the CDF is 92.5%, however the Top-1 accuracy of Modified SIFT is 57.0%, and the combination of the CDF and the modified SIFT achieves only 89.0%. Compared with the sole CDF and the modified SIFT, the combination of both features dramatically degrades the performance. It indicates that the modified SIFT has adverse effect on the combination which leads to performance degradation. The reason for the poor performance of the modified SIFT is that there are too few characters in the mail address image. Essentially, the modified SIFT is the bag of visual words model based feature. Without sufficient characters, we only extract few keypoints from the address image. This will cause the frequency histogram of visual words is not able to represent the characteristics of the writer. Hence, the modified SIFT is not useful when the handwriting image contains few characters.

4. Conclusions

Biometrics are widely deployed to ensure the personal and public security in our daily life. Writer identification as a special case of behavioral biometrics, is facing great challenges and chances. We concentrate on the particular issue of off-line text-independent Chinese writer identification, and propose a new method based on the contour-directional feature and the modified SIFT. Experiments on HIT-MW dataset demonstrate the effectiveness of the proposed method. The modified SIFT is insensitive to the aspect ratio, and the contour-directional feature can capture the local details of the stroke structures. The new concatenated feature is capable of reflecting the characteristics of handwriting appropriately. Experiments on LPAIS dataset also show the weakness of the modified SIFT when the number of characters is few. Without enough keypoints extracted from the handwriting image, the bag of visual words model based feature cannot provide a proper feature representation for the characteristics of individual writing style. In the feature, we will pay more attention to explore how to deal with the situation when the handwriting image contains few characters.

References

1. A. K. Jain, A. Ross, and S. Pankanti, Biometrics: a tool for information security, *IEEE Transactions on Information Forensics and Security.* **1**(2), 125–143 (2006).
2. R. Plamondon and S. N. Srihari, Online and off-line handwriting recognition:

a comprehensive survey, *IEEE Transactions on Pattern Analysis and Machine Intelligence*. **22**(1), 63–84 (2000).
3. M. L. Bulacu. *Statistical pattern recognition for automatic writer identification and verification*. Ph.d. thesis, University of Groningen (2007).
4. X. Y. Zhang and C. L. Liu. Style transfer matrix learning for writer adaptation. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, pp. 393–400 (2011).
5. M. L. Bulacu and L. Schomaker, Text-independent writer identification and verification using textural and allographic features, *IEEE Transactions on Pattern Analysis and Machine Intelligence*. **29**(4), 701–717 (2007).
6. D. Impedovo and G. Pirlo, Automatic signature verification: the state of the art, *IEEE Transactions on System, Man and Cybernetics, Part C: Applications and Reviews*. **38**(5), 609–635 (2008).
7. R. Plamondon and G. Lorette, Automatic signature verification and writer identification - the state of the art, *Pattern Recognition*. **22**(2), 107–131 (1989).
8. Y. Zhu, T. N. Tan, and Y. H. Wang. Biometric personal identification based on handwriting. In *Proceedings of the International Conference on Pattern Recognition*, pp. 797–800 (2000).
9. C. Shen, X. G. Ruan, and T. L. Mao. Writer identification using gabor wavelet. In *Proceedings of the World Congress on Intelligent Control and Automation*, pp. 2061–2064 (2002).
10. Z. Y. He and Y. Y. Tang. Chinese handwriting-based writer identification by texture analysis. In *Proceedings of the International Conference on Machine Learning and Cybernetics*, pp. 3488–3491 (2004).
11. Z. Y. He, X. G. You, and Y. Y. Tang, Writer identification of chinese handwriting documents using hidden markov tree model, *Pattern Recognition*. **41**(4), 1295–1307 (2008).
12. Z. Y. He, X. G. You, and Y. Y. Tang, Writer identification using global wavelet- based features, *Neurocomputing*. **71**(10), 1832–1841 (2008).
13. J. J. Zhang, Z. Y. He, Y. M. Cheung, and X. G. You. Writer identification using a hybrid method combining gabor wavelet and mesh fractal dimension. In *Proceedings of the International Conference on Intelligent Data Engineering and Automated Learning*, pp. 535–542 (2009).
14. X. Li and X. Q. Ding. Writer identification of chinese handwriting using grid microstructure feature. In *Proceedings of the International Conference on Biometrics*, pp. 1230–1239 (2009).
15. G. Louloudis, N. Stamatopoulos, and B. Gatos. ICDAR 2011 writer identification contest. In *Proceedings of the International Conference on Document Analysis and Recognition*, pp. 1475–1479 (2011).
16. L. Xu, X. Q. Ding, L. Peng, and X. Li. An improved method based on weighted grid micro-structure feature for text-independent writer recognition. In *Proceedings of the International Conference on Document Analysis and Recognition*, pp. 638–642 (2011).
17. J. Wen, B. Fang, J. L. Chen, Y. Y. Tang, and H. X. Chen, Fragmented edge structure coding for chinese writer identification, *Neurocomputing*. **86**, 45–51

(2012).
18. Y. J. Hu, W. M. Yang, and Y. B. Chen. Bag of features approach for offline text-independent chinese writer identification. In *Proceedings of the International Conference on Image Processing*, pp. 2609–2613 (2014).
19. M. L. Bulacu, L. Schomaker, and L. Vuurpijl. Writer identification using edge-based directional features. In *Proceedings of the International Conference on Document Analysis and Recognition*, pp. 937–941 (2003).
20. D. G. Lowe, Distinctive image features from scale-invariant keypoints, *International journal of computer vision*. **60**(2), 91–110 (2004).
21. T. H. Su, T. W. Zhang, and D. J. Guan, Corpus-based HIT-MW database for offline recognition of general purpose chinese handwritten text, *International Journal on Document Analysis Recognition*. **10**(1), 27–38 (2007).
22. L. Van Der Maaten and E. Postma. Improving automatic writer identification. In *Proceedings of the Belgium-Netherlands Conference on Artificial Intelligence*, pp. 260–266 (2005).