Chinese Writer Identification Using Contour-directional Feature and Character Pair Similarity Measurement

Yu-Jie Xiong, Yue Lu
Shanghai Key Laboratory of Multidimensional Information Processing
Department of Computer Science and Technology
East China Normal University, Shanghai 200241, China
E-mail: xiong@stu.ecnu.edu.cn; ylu@cs.ecnu.edu.cn

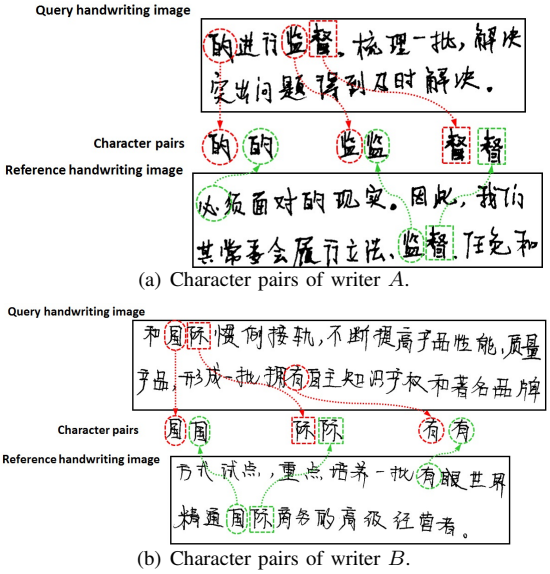
***Abstract*—The key issue of Chinese writer identification is the uncertainty of the text content in the query and reference handwriting images. We propose a method for Chinese writer identification using Contour-directional Feature (CDF) and Character Pair Similarity Measurement (CPSM). CDFs are extracted from the query and reference handwriting images and are used to calculate the text-independent similarity between the query and reference handwriting images. Meanwhile, characters appearing in both the query and reference handwriting images are also utilized to measure the similarity of character pairs. The text-independent similarity and the similarity of character pairs are fused to the final similarity between the query and reference handwriting images. The proposed method is evaluated on two public datasets. The best Top-1 identification accuracy on the HIT-MW and CASIA-2.1 dataset reaches 96.7% and 97.9% respectively, which outperforms other previous approaches.**

***Keywords*-Chinese writer identification, contour-directional feature, character pair similarity measurement, keypoint matching, similarity fusion.**

I. Introduction

Writer identification is to identify the authorship of handwritings. It can be generally classified into two categories: text-independent and text-dependent [1]. Text-independent methods analyze the writing style of handwriting and require sufficient handwritten text to extract robust features. Thus, the minimal amount of handwritten text is of crucial importance. On the other hand, text-dependent methods require that the text content of the query and reference handwriting images are totally the same and apply direct comparison bwtween the text content appearing in both the query and reference handwriting images.

As for Chinese writer identification, Zhu et al. [2] adopted the Gabor filtering technique for text-independent writer identification. Wang et al. [3] extracted directional element features from a single character to identify the writer. Li et al. [4] proposed a writer retrieval system based on text-sensitive writer identification. Li and Ding [5] proposed Grid Microstructure Feature (GMF) to describe the characteristics of Chinese handwriting. Xu et al. [6] proposed a weighted feature matching method for writer identification based on the inner and inter class variances. Hu et al. [7] employed



(a) Character pairs of writer A.

(b) Character pairs of writer B.

Figure 1: Examples of characters appearing in the query and reference handwriting images.

SIFT descriptors and presented two coding strategies for feature coding. Wu et al. [8] utilized descriptors, scale and orientation of the region-based SIFT for writer identification.

By investigating the text content of both the query and reference handwriting images, we find that the text content of the query and reference handwritings are not totally different in most cases. Some high-frequency characters appear in both the query and reference handwritings. Fig. 1 presents two handwriting images (one query and one reference) of writer A and two handwriting images of writer B. There are three character pairs (‘的’, ‘监’, ‘督’) that appear in both the query and reference handwriting images of writer A in Fig. 1(a) and three character pairs (‘ 国’, ‘际’, ‘有’) that appear in both the query and reference handwriting images of writer B in Fig. 1(b). The characters pairs appearing in both the query and reference handwriting images are helpful to identify the writer of the query handwriting. Thus, our motivation is to utilize the characteristics of character pairs

2379-2140/17 $31.00 © 2017 IEEE
DOI 10.1109/ICDAR.2017.28

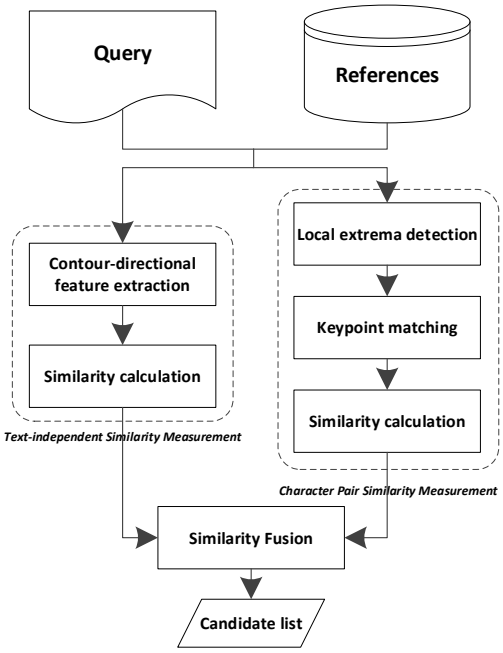


Figure 2: The flowchart of the proposed method.

to improve text-independent Chinese writer identification. The flowchart of the proposed method method is given in Fig. 2. CDFs are extracted to represent the writing style of handwriting. Then, the weighted Chi-squared metric is used to measure the text-independent similarity between the query and reference handwriting images. Meanwhile, the proposed CPSM are utilized to calculate the similarity of character pairs. Local extrema detection is employed to find keypoints from the characters appearing in both the query and reference handwriting images. Then, the Euclidean distance of SIFT descriptors of the keypoints are used for keypoint matching. Geometric constraints are utilized to help eliminate false matches. After that, SIFT descriptors of these matching keypoint pairs are used to calculate the similarity using the Manhattan metric. Finally, the text-independent similarities of the query and reference handwriting images are fused with the corresponding similarities of character pairs to rank the list of the reference handwriting images from the most similar to the least similar. The writer of the first reference handwriting image in the reordered list is considered as the most possible writer of the query handwriting image.

The remainder of the paper is organized as follows: we present the details of the proposed method in Section 2 and Section 3. The experimental results are given in Section 4. We conclude the work in Section 5.

II. Text-independent similarity measurement

Text-independent similarity measurement consists of two stages. The first stage is to extract CDFs from the query and reference handwriting images and the second stage is to calculate the similarities between the query handwriting image and the reference handwriting images using the obtained CDFs. Xiong et al. [9] proposed CDF for English and Geek writer identification and achieved satisfying results. In this paper, we extend CDF to Chinese writer identification.

A. Contour-directional feature extraction

Contour-directional feature is an occurrence histogram of particular edge pixel pairs. It can be represented as $(a_1, a_2, \cdots, a_C)$ and $\sum_{i=1}^{C} a_i = 1$, where C is the dimension of the vector and each element x_i $(1 \leq i \leq C)$ of the vector is the probability of the occurrence of a category of particular pixel pairs that have the same directional index.

The frequency of the occurrence of all particular edge pixel pairs in every local grid are recorded according to their directional indices and normalized into an occurrence histogram of particular edge pixel pairs. To obtain edge pixel pairs, the contour image is divided into a number of grids, and the center of each grid is an edge pixel. Assume that there is a grid of $(2W+1) \times (2W+1)$ and its center pixel is P, where W is the chessboard distance between P and the outermost pixels in the grid. The binary function $V(\bullet)$ is defined as

$$V(\bullet) = \begin{cases} 1, & The\ pixel\ is\ an\ edge\ pixel, \\ 0, & Otherwise, \end{cases} \tag{1}$$

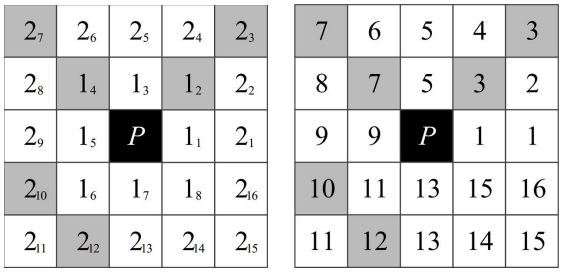
and used to indicate whether the pixel is a edge pixel or not. The rest of pixels in the gird are denotes as w_i, where $w = 1, 2, \cdots, W$ is the chessboard distance of the pixel to P and i $(1 \leq i \leq 8 * w)$ is the index to distinguish pixels that have the same chessboard distance to P. There are $8*w$ pixels surround the center P with w, and they are assigned from w_1 to w_{8*w}. A binary function $EP(\varepsilon, \zeta)$ is defined as

$$EP(\varepsilon, \zeta) = \begin{cases} 1, & V(\eta) = 0, V(\varepsilon) = V(\zeta) = 1 \\ 0, & Otherwise, \end{cases} \tag{2}$$

where $\varepsilon = w_d$, $\zeta = w_e$, $\eta = w_f$ and $1 \leq d < f < e \leq 8 * w_1$. It is used to determine whether the pixel pairs (ε, ζ) is a edge pixel pairs or not. The direction $Dir(\bullet)$ of the pixel is defined the angle between the line from P to the pixel and the horizontal line. Assume that the directions of pixels appearing in the grid is denoted as $Dir(\bullet)_1$, $Dir(\bullet)_2$, $\cdots$, $Dir(\bullet)_J$, where J is the number of different directions of the pixel appearing the in the grid. According to this assumption, the direction $Dir(w_i)$ of each pixel w_i can be denoted as

$$Dir(w_i) = Dir(\bullet)_h \ (1 \leq h \leq J). \tag{3}$$

Afterwards, a series of counters $\{CT(j,k) = 0 | 1 \leq j < k \leq J\}$ are initialized. When the gird is moved to the center



(a) Find edge pixel pairs. (b) Update the indices of pixels.

Figure 3: An example of the extraction of CDF.

of a edge pixel and there is a pixel pair (ε, ζ) meets the condition of $EP(\varepsilon, \zeta) = 1$,

$$CT(j,k) = \begin{cases} CT(j,k)+1, & Dir(\varepsilon) = Dir(\bullet)_j, \\ & Dir(\zeta) = Dir(\bullet)_k, \\ CT(j,k), & Otherwise, \end{cases} \quad (4)$$

the corresponding counter adds one count and the rest of counters remain the same. After each local grid of the handwriting image has been traversed, all edge pixel pairs appearing in the handwriting image are also recorded. The sum T of all counters $CT(j,k)$ is calculated as

$$T = \sum_{j,k}^{1 \le j < k \le J} CT(j,k). \quad (5)$$

The normalized occurrence histogram of edge pixel pairs $(\frac{CT(1,2)}{T}, \frac{CT(1,3)}{T}, \cdots, \frac{CT(J-1,J)}{T})$ is regarded as the contour-directional feature vector. Fig. 3 shows an example of the extraction of CDF. There are four edge pixel pairs $((1_2, 1_4), (2_3, 2_7), (2_7, 2_{10}), (2_{10}, 2_{12}))$ in Fig. 3(a). According to the corresponding directional indices in Fig. 3(b), they are recorded in three counters ($CT(3,7) = 2$, $CT(7,10) = 1$, $CT(10,12) = 1$). While, GMF [5] does not consider the directional information of edge pixel pairs. According to the definition of GMF, $(1_2, 1_4)$ and $(2_3, 2_7)$ are treated as two different kinds of edge pixel pairs. Consequently, four edge pixel pairs should be recorded in four different counters.

B. Similarity calculation

After contour-directional features are extracted from the query and reference handwriting images, the weighted Chi-squared metric is used to calculate the distance between them. Assume that there are L reference handwriting images, and the query handwriting image and reference handwriting images are denoted as Q and R_l $(1 \le l \le L)$, respectively. Let CDF_Q $(a_1, a_2, ..., a_C)$ and CDF_R^l $(b_1^l, b_2^l, ..., b_C^l)$ denote their contour-directional features. The distance D_C between CDF_Q and CDF_R^l is computed by:

$$D_C = \sum_{n=1}^{N} \frac{(a_n - b_n^i)^2}{(a_n + b_j^n) * \sigma_n}, \quad (6)$$

where $\sigma_n = \sqrt{\frac{1}{L-1}\sum_{l=1}^{L}(b_n^l - \mu_n)^2}$, and $\mu_n = \frac{1}{L}\sum_{l=1}^{L} b_n^l$.

III. CHARACTER PAIR SIMILARITY MEASUREMENT

In this study, the character pairs of the same text content appearing in both the query and reference handwriting images, are compared to evaluate their similarity. It contains three steps: local extrema detection, keypoint matching, and similarity calculation.

A. Local extrema detection

The local extrema detection is a part of the standard SIFT. The main idea of the local extrema detection is to find keypoints in all possible scale. It is used to find keypoints of characters appearing in both the query and reference handwriting images. Readers can find more details about local extrema detection in the Ref. [10].

B. keypoint matching

After the local extrema detection, keypoints of characters appearing in both the query and reference handwriting images are obtained. Not all keypoints are used to calculate the similarity of character pairs. Only the dual matched keypoint pair will be used. Traditional method of keypoint matching is measured by the Euclidean distance of SIFT descriptors. However, the SIFT descriptors do not contain any position information. This drawback may leads to false matches. Therefore, we utilize geometric constraints to eliminate the false matches. The process can be summarized as follows:

Step 1. Calculate candidate matching keypoint pairs based on the Euclidean distance.

C_A and C_B is a character pair appearing in both the query and reference handwriting images, respectively. There are Num_A keypoints in character C_A, and Num_B keypoints in character C_B, respectively. $O = \{o_a | 1 \le a \le Num_A\}$ is a set of keypoints of C_A in the query handwriting image, and $S = \{s_b | 1 \le b \le Num_B\}$ is a set of keypoints of C_B in the reference handwriting image. We calculate the Euclidean distance between SIFT descriptors of keypoints in O and that of keypoints in S. For each o_a, we select the first five nearest keypoints in S to build the set V of candidate matching keypoint pairs.

Step 2. Create the graph G based on geometric constraints.

Assume that there are two candidate matching keypoint pairs (o_i, s_k) and (o_j, s_l), where x_{o_i}, x_{o_j}, x_{s_k}, x_{s_l} and y_{o_i}, y_{o_j}, y_{s_k}, y_{s_l} are the horizontal and vertical axis coordinates of keypoints o_i, o_j, s_k, and s_l, respectively. Geometric constraints between them are defined as:

$$\begin{aligned} \left|(x_{o_i} - x_{o_j}) - (x_{s_k} - x_{s_l})\right| &< \Delta_1 \times Width_A, \\ \left|(y_{o_i} - y_{o_j}) - (y_{s_k} - y_{s_l})\right| &< \Delta_2 \times Height_A. \end{aligned} \quad (7)$$

Where Δ_1 and Δ_2 are parameters to adjust geometric constraints. $Width_A$ and $Height_A$ are the width and height of the character C_A in the query handwriting image.

(a) Characters of different writers. (b) Characters of the same writer.

Figure 4: Matching keypoint pairs without geometric constraints.

(a) Characters of different writers. (b) Characters of the same writer.

Figure 5: Matching keypoint pairs with geometric constraints.

We use candidate matching keypoint pairs in V to create a graph G on the basis of geometric constraints. The keypoint pair in V is denoted as a vertex in G. If keypoint pairs (o_i, s_k) and (o_j, s_l) satisfy geometric constraints, then there is an edge between their corresponding vertexes in G. Otherwise their corresponding vertexes are not connected.

Step 3. Find the maximal clique in G.

We use the vertex to denote the candidate matching keypoint pair, and the edge of vertexes to denote the relationship between keypoint pairs. Thus, the matching problem is turned into the maximum clique problem of the graph G. A simple yet effective method is used to achieve this goal. At first, we calculate the degree of each vertex in G and initialize a new set F. Then, we find the vertex with the maximum degree and move it into $F = \emptyset$. After that, the vertexes in G which are not connected to the vertex in F are removed from G and abandoned. This process is repeated until the stop condition $G = \emptyset$. The vertexes in the set F can build the maximal clique of G, and keypoint pairs of corresponding vertexes are the final matching keypoint pairs.

Fig. 4 and Fig. 5 show the influence of geometric constraints for keypoint matching. There are 4 matching keypoint pairs in Fig. 4(a) and 10 matching keypoint pairs in Fig. 4(b), respectively. Both of them contain 2 false matches. As shown in Fig. 5(a) and Fig. 5(b), false matches are eliminated with geometric constraints and the remaining keypoint pairs are matched correctly. Fig. 5 also shows that the matching keypoint pairs of characters from the same writer (Fig. 5(b)) are more than that of characters from different writers (Fig. 5(a)). It helps us to determine whether the query handwriting and reference handwriting images are created by the same writer or not.

C. Similarity calculation

All character pairs appearing in both the query and reference handwriting images are used for similarity calculation. Assume that the character pair appearing in both the query handwriting image Q and reference handwriting image R is denoted as (C_Q^e, C_R^e) and the number of character pairs is denoted as E, where $1 \leq e \leq E$. The SIFT descriptors of matching keypoint pairs in (C_Q^e, C_R^e) are denoted as $((x_1^h, x_2^h, ..., x_{128}^h), (y_1^h, y_2^h, ..., y_{128}^h))$ and the number of matching keypoint pairs is denoted as H_i, where $1 \leq h \leq H_e$. The Manhattan metric is adopted to measure the similarity d_e between C_Q^e and C_R^e:

$$d_e = \frac{1}{H_e} \sum_{h=1}^{H_e} \sum_{q=1}^{128} \left| x_q^h - y_q^h \right|. \tag{8}$$

And the average distance of all E character pairs is defined as the similarity between Q and R:

$$D_S = \frac{1}{E} \sum_{e=1}^{E} d_e. \tag{9}$$

D. Similarity fusion

When two similarities between the query and reference handwriting images are calculated using CDF and CPSM, we normalize D_C and D_S into interval [0, 1]. Then, we summarize them together to measure the final similarity between Q and R:

$$D_Q^R = \delta * D_S + (1 - \delta) * D_C, \tag{10}$$

where $0 < \delta < 1$ is the weight parameter to balance the contribution of CDF and CPSM.

IV. Experimental results

We evaluate the proposed method on the HIT-MW [11] and CASIA-2.1 [12] datasets. In our experiments, as done in [5], only the first page of 240 writers is used. Each page is segmented into two commensurate parts. The CASIA-2.1 dataset contains two sub-datasets, we use the larger one which contains handwritings of 240 writers.

Both datasets are divided into the query and reference set, and every writer only has one image in each set. Given a query handwriting image Q, the system sorts all images in the reference set based on the their similarities compared with Q. Ideally, the reference handwriting image with the minimum distance should be created by the same writer of Q. Ranking list (Top-N) is used to measure the performance of the proposed method. For the Top-N criterion, a correct hit is accumulated when at least one handwriting in the first N place of the ranking list is created by the correct writer. In our experiments, we use the identification accuracy of Top-1, Top-5, and Top-10. The parameters Δ_1 and Δ_2 are set as 0.4 and 0.3, empirically. The ground truth of the HIT-MW and CASIA-2.1 dataset is utilized to find characters appearing in both the query and reference handwriting images.

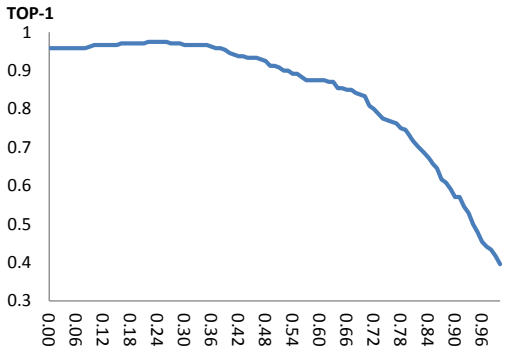


Figure 6: The Top-1 accuracy with different δ on the HIT-MW dataset.

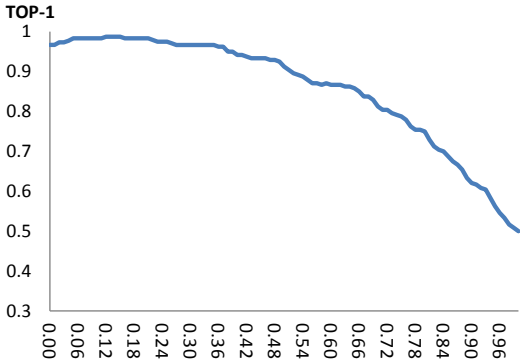


Figure 7: The Top-1 accuracy with different δ on the CASIA-2.1 dataset.

A. *Parameter selection*

The size of grid has influence on the effectiveness of CDF. The local structure information is fragmentized when the size of grid is too small, while the stroke information is rough when the grid size is too large. A suitable selection of the grid size is related to the character size of the handwriting samples. We use different sizes of grid to extract CDF from the handwriting images and evaluate the performance of obtained CDFs. Tab. I gives the Top-1 accuracy of CDF that are extracted by different grid sizes on both datasets. When the size of grid is 15×15, the identification accuracy is the highest, hence we choose 15×15 as the best grid size for all the experiments. In fact, the height of the characters samples is about 40 to 90 pixels. The grid of 15×15 is able to capture both the stroke and local structure information of the character at the same time.

Table I: The Top-1 accuracy of CDF extracted by different grid sizes.

Dataset \ Size	9×9	11×11	13×13	15×15	17×17
HIT-MW	93.8%	94.6%	95.0%	**95.8%**	95.4%
CASIA-2.1	94.2%	95.4%	95.8%	**97.1%**	96.3%

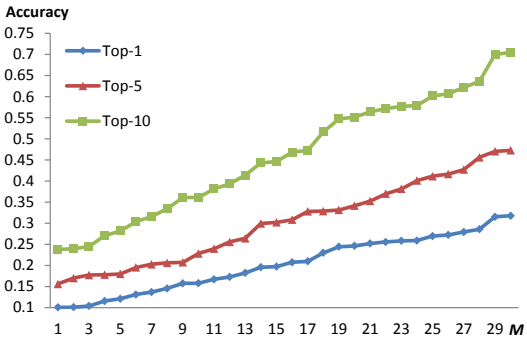


Figure 8: The performance of the CPSM with different character pairs on the HIT-MW dataset.

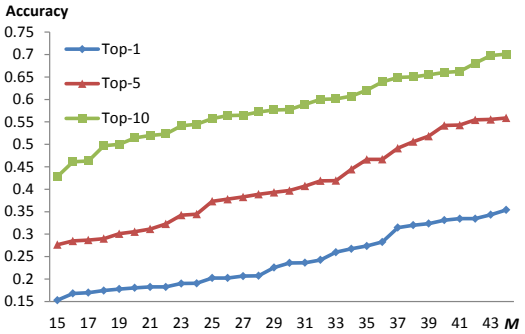


Figure 9: The performance of the CPSM with different character pairs on the CASIA-2.1 dataset.

We carry out the experiment to find the optimal weight parameter δ of two datasets. The value of δ is selected from 0 to 1 incrementally. For each value, we calculate the Top-1 accuracy to investigate the effect of δ on identification performance. Fig. 6 and Fig. 7 show the Top-1 accuracy with different δ on two datasets, respectively. We can see that the optimal weight parameter δ on the HIT-MW dataset is 0.25, while the corresponding value on the CASIA-2.1 dataset is 0.13. In consideration of the fact that the average amount of characters contained in the images of CASIA-2.1 dataset is twice than that of characters contained in the images of HIT-MW dataset, the optimal δ may be related to the amount of characters contained in the handwriting image. With more characters in the handwriting, the extracted CDF is more effective to describe the writing style of handwriting, and the character pairs information has less contribution to improve the identification performance.

B. *The amount of character pairs and its influence on CPSM*

To further investigate the influence of characters appearing both the query and reference handwriting images, we perform the experiment to study the relationship of the amount of character pairs and the identification accuracy of CPSM. We only use the first M character pairs in both the query

Table II: The accuracy of different methods on the HIT-MW dataset.

Top-N \ Feature	Li [5]	Wu [8]	CDF [9]	CPSM	CDF+CPSM
Top-1	95.0%	95.4%	95.8%	39.6%	**96.7%**
Top-5	98.3%	98.8%	98.8%	53.8%	**99.2%**
Top-10	98.8%	**99.2%**	**99.2%**	76.3%	**99.2%**

Table III: The accuracy of different methods on the CASIA-2.1 dataset.

Top-N \ Feature	Li [5]	Hu [7]	CDF [9]	CPSM	CDF+CPSM
Top-1	90.0%	96.3%	97.1%	50.0%	**97.9%**
Top-5	NULL	NULL	98.8%	67.9%	**99.2%**
Top-10	97.1%	**99.6%**	**99.6%**	78.3%	**99.6%**

and reference handwriting images for CPSM. According to distributions of the amount of character pairs on two datasets, the range of M on the HIT-MW dataset is 1 to 30, and the range of M on the CASIA-2.1 dataset is 15 to 44. The Top-N accuracy of CPSM with different character pairs on two datasets are shown on Fig. 8 and Fig. 9, respectively. It is obvious that the more character pairs appearing in both the query and reference handwriting images are utilized, the better performance is achieved.

C. Comparison of the proposed method with others

We compare the proposed method with previous text-independent Chinese writer identification methods. Tab. II and Tab. III show the performance of different methods on two datasets, respectively. The accuracy of CDF is better than that of other previous methods on the both datasets. It demonstrates that CDF keeps more local structural information, not only the relationship of adjacent strokes but also the direction of pixel pairs. Due to the fact that CPSM is not totally text-independent, the identification accuracy of CPSM is far below that of text-independent methods. But CPSM characterize the handwriting from a different way and as a complementarity of text-independent features, it improves the overall identification performance.

V. Conclusion

We propose a method for Chinese writer identification. The proposed method are not only able to applied to text-independent writer identification but also utilizes the character pairs to improve the identification performance. In order to exploit information of character pairs, we propose the Character Pair Similarity Measurement (CPSM) to calculate the similarity of character pairs appearing in both the query and reference handwriting images. Experimental results show that CPSM can enhance the identification performance of text-independent features, and our method outperforms other previous approaches on two datasets.

Acknowledgment

This work is jointly supported by the Science and Technology Commission of Shanghai Municipality under research grants 14511105500 and 14DZ2260800.

References

[1] R. Plamondon and G. Lorette, "Automatic signature verification and writer identification - the state of the art," *Pattern Recognition*, vol. 22, no. 2, pp. 107–131, 1989.

[2] Y. Zhu, T. N. Tan, and Y. H. Wang, "Biometric personal identification based on handwriting," in *Proceedings of the International Conference on Pattern Recognition*, 2000, pp. 797–800.

[3] X. L. Wang, X. Q. Ding, and H. L. Liu, "Writer identification using directional element features and linear transform," in *Proceedings of the International Conference on Document Analysis and Recognition*, 2003, pp. 942–945.

[4] X. Li, X. L. Wang, and X. Q. Ding, "An off-line Chinese writer retrieval system based on text-sensitive writer identification," in *Proceedings of the International Conference on Pattern Recognition*, 2006, pp. 517–520.

[5] X. Li and X. Q. Ding, "Writer identification of Chinese handwriting using grid microstructure feature," in *Proceedings of the International Conference on Biometrics*, 2009, pp. 1230–1239.

[6] 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*, 2011, pp. 638–642.

[7] 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*, 2014, pp. 2609–2613.

[8] X. Q. Wu, Y. B. Tang, and W. Bu, "Off-line text-independent writer identification based on scale invariant feature transform," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 3, pp. 526–536, 2014.

[9] Y.-J. Xiong, Y. Wen, P. S. Wang, and Y. Lu, "Text-independent writer identification using SIFT descriptor and contour-directional feature," in *Proceedings of the International Conference on Document Analysis and Recognition*, 2015, Conference Proceedings, pp. 91–95.

[10] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[11] T. H. Su, T. W. Zhang, and D. J. Guan, "Corpus-based HIT-MW database for off-line recognition of general purpose Chinese handwritten text," *International Journal on Document Analysis Recognition*, vol. 10, no. 1, pp. 27–38, 2007.

[12] C. L. Liu, F. Yin, D. H. Wang, and Q. F. Wang, "CASIA online and off-line Chinese handwriting databases," in *Proceedings of the International Conference on Document Analysis and Recognition*, 2011, pp. 37–41.