I.J. Intelligent Systems and Applications, **2011, 2, 1-8**
Published Online March 2011 in MECS (http://www.mecs-press.org/)

A Stroke Shape and Structure Based Approach for Off-line Chinese Handwriting Identification

Jun Tan, Jian-Huang Lai, Chang-Dong Wang
School of Information Science and Technology, Sun Yat-sen University, Guangzhou, P. R. China.
Email: mcstj@mail.sysu.edu.cn, stsljh@mail.sysu.edu.cn, mc04wchd@mail2.sysu.edu.cn

Ming-Shuai Feng
Public Security of Guangdong Province, Guangzhou, P. R. China.

Abstract—**Handwriting identification is a technique of automatic person identification based on the personal handwriting. It is a hot research topic in the field of pattern recognition due to its indispensible role in the biometric individual identification. Although many approaches have emerged, recent research has shown that off-line Chinese handwriting identification remains a challenge problem. In this paper, we propose a novel method for off-line Chinese handwriting identification based on stroke shapes and structures. To extract the features embedded in Chinese handwriting characters, two special structures have been explored according to the trait of Chinese handwriting characters. These two structures are the bounding rectangle and the TBLR quadrilateral. Sixteen features are extracted from the two structures, which are used to compute the unadjusted similarity, and the other four commonly used features are also computed to adjust the similarity adaptively. The final identification is performed on the similarity. Experimental results on the SYSU and HanjaDB1 databases have validated the effectiveness of the proposed method.**

Index Terms—**handwriting identification, off-line, Chinese character, stroke, mathematical morphology, feature extraction**

I. Introduction

As one of the most important methods in the biometric individual identification, handwriting identification has been widely used in the fields of bank check [1], forensic [2], historic document analysis [3], archaeology [4], identifying personality [5], etc. It is a hot research topic with the aim of automatically identifying a person based on the personal handwriting. Many approaches have been developed [1]-[6]. According to the different input methods, handwriting identification is commonly classified into on-line and off-line. The former assumes that a transducer device is used to capture the writing information such as time order and dynamics when a writer is writing the characters. Off-line technique, however, only deals with handwriting images scanned into computer, leading to the lost of dynamic information. Therefore, compared with its on-line counterpart, off-line handwriting identification is a rather challenging problem.

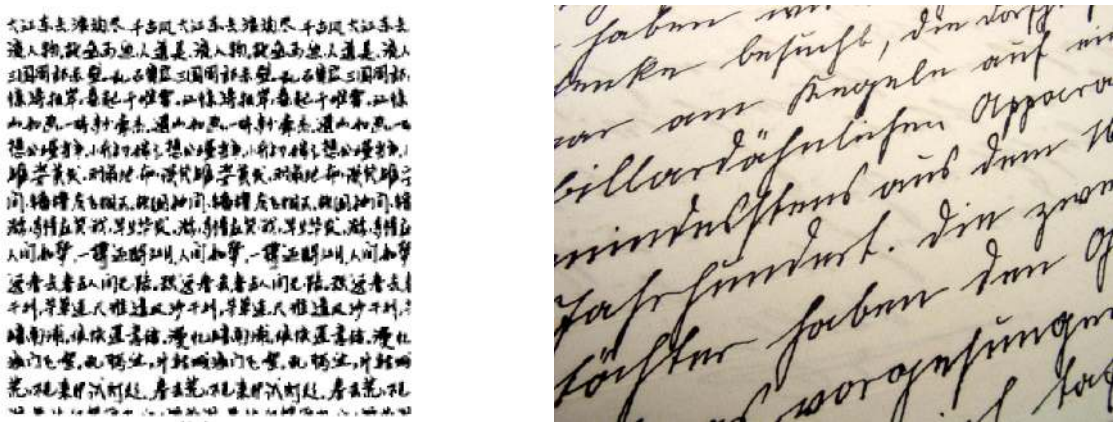
Chinese characters are ideographic in nature, which contain at least 50000 characters. However, only 6000 of them are commonly used and they have a wide range of complexity. Chinese characters can be expressed in at least two common styles, such as in block or in cursive. In block style, there is an average of 810 strokes. Meanwhile there are more strokes in cursive style. According to [17], in Chinese characters, the complication structures are mostly affected by multi stokes of each character. Additionally, as shown in Figure 1, the stroke shapes and structures of Chinese characters are quite different from those of other languages such as English, which makes it more difficult to identify Chinese handwriting [6]. The approaches proposed for English handwriting identification is no longer suitable for the case of Chinese handwritings [2] [3] [11]. In this paper, we mainly focus on off-line Chinese handwriting identification, and propose a novel method for extracting a set of twenty features based on two newly proposed special structures according to the trait of Chinese handwriting characters.

A. Related Work

The process of handwriting identification consists of three main parts: preprocessing, feature extraction and classification (or matching). The feature extraction and matching are the two major topics in the literature of handwriting identification.

Features such as texture, edge, contour and character shape have been widely studied recently. Several researchers [6]–[8] proposed to take the handwriting as an image containing special texture, and therefore regarded the handwriting identification as the texture identification. Among them, Zhu [7] and He [6] adopted 2-D Gabor filtering to extract the texture features, while Chen et al. [8] used the Fourier transform. To reduce the computational cost suffered by 2-D Gabor filters, He et al. [9] further introduced a contourlet method to handwriting identification. In [10], edge-based directional probability distributions were used as features; meanwhile character-shape (allograph) is another type of effective feature [2]. In [15], the feature vector was derived by

This work was supported by the Science and Technology Program of Guangdong Province under Grant 2007B030603003.



(a) Sample of Chinese handwriting

(b) Sample of English handwriting

Figure 1. Comparing the Chinese and English handwritings.

morphologically processing the horizontal profiles of the words, where the projections were derived and processed in segments to increase the discriminating power.

The widely used classifiers at least include Hidden Markov Model (HMM) [11] [12], weighted Euclidean distance (WED) classifier [6]-[8], Bayesian model [2] [15], likelihood ranking [3], etc. In [11], a Hidden Markov Model (HMM) based recognizer was built for each writer and trained on text lines written by the corresponding writer. For eliminating the disturbance caused by unexpected noise, which may "break" the normal transmission of states in the observation sequences, Ko et al. [12] suggested using a leave-one-out-training and testing strategy to make HMMs more robust. For matching singleton non-sequential features such as texture, edge and contour, the weighted Euclidean distance (WED) [6]-[8] has been shown to be effective by the experiments. In [15], both Bayesian classifiers and neural networks were used as the classifiers.

Focusing on Chinese handwriting identification, some particular methods have emerged recently [6] [7] [16] [18]. In [18], Li and Ding proposed a histogram-based feature, called grid microstructure feature which is extracted from the edge image of the scanned images. In [7], Zhu et al. took the handwriting as an image containing some special texture and Chinese handwriting identification is regarded as texture identification. Similarly, the texture based approach is also used in [6], where both text-independent and text-dependent methods have been introduced. A contourlet-based method was proposed in [9].

B. Our Approach

In this paper, a novel method is proposed to extract a set of twenty features based on stroke shape and structure. Two special structures of the Chinese handwriting character are explored, including the bounding rectangle and TBLR quadrilateral. From the bounding rectangle, nine features are extracted; while another seven features are computed based on the TBLR quadrilateral. These sixteen features are used together to compute the unadjusted similarity. Then another four commonly used features are computed to adaptively adjust the similarity that is already evaluated. The identification is finally performed on the adjusted similarity. Experiments on the SYSU and HanjaDB1 databases are conducted to compare the proposed method with two algorithms. Comparison results have shown the effectiveness of the proposed method.

Some of the results in this paper were first presented in [14]. In this paper, we present more technique details concerning the rationality of the proposed approach and report more experimental results to confirm the

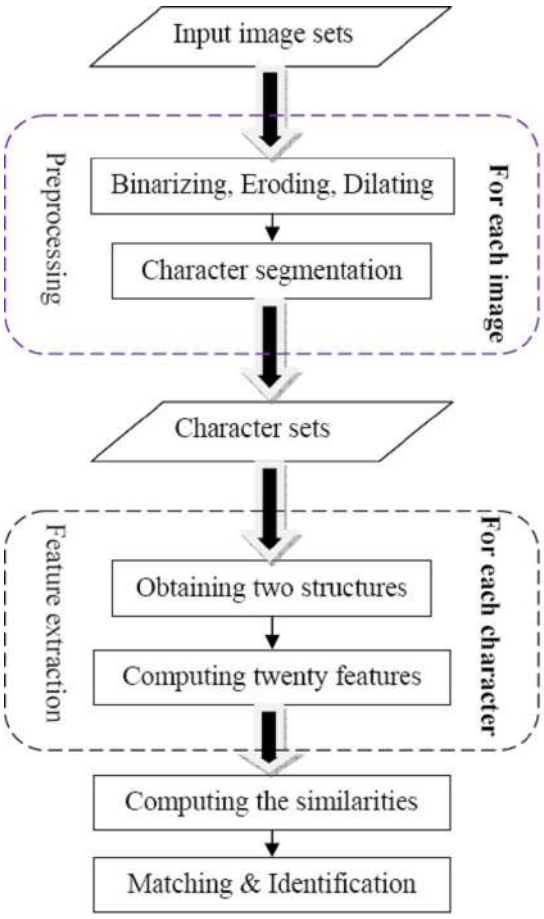


Figure 2. The flowchart of the proposed approach.

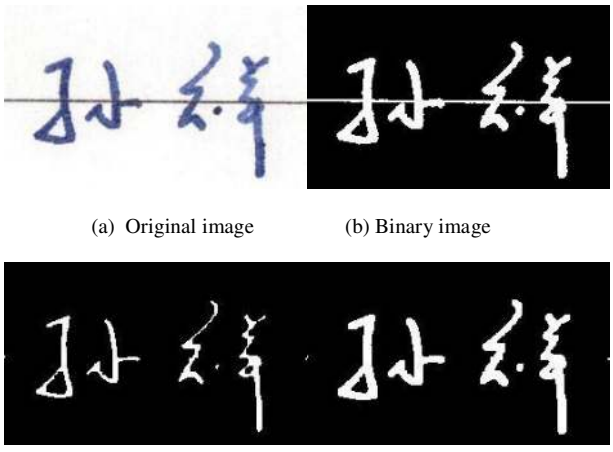
effectiveness of the proposed approach. Figure 2 demonstrate the flowchart of the proposed approach.

The remaining of this paper is organized as follows. In section II, we introduce the preprocessing phase of our approach, which uses mathematical morphology for removing the cluttered and thin background. Section III describes the feature extraction, where two special structures are proposed for facilitating the extraction of twenty features. The matching phase is described in section IV. Section V reports the experimental results. The conclusions are drawn in section VI.

II. Preprocessing

In the real-world applications, the images obtained are usually with cluttered background, even noises. Additionally, the scanned characters may be of different sizes, and have different spaces between text lines. Therefore, a preprocessing is often required [7] [9] [15]. The Common steps used for pre-processing are as follows. The noises are firstly removed from the handwriting image; secondly, the text line is located and the single character is obtained using projection; thirdly, each character is normalized into a same size. Since for Chinese handwriting identification, there often exist horizontal background lines that are much thinner than the foreground stroke, as shown in Figure 3(a). In this paper, we use mathematical morphology for removing the cluttered and thin background, which is also a commonly used approach [7] [15].

Three main steps in our preprocessing phase include binarizing, eroding, and dilating. Let A be a binary



(a) Original image (b) Binary image

(c) Eroded image (d) Dilated image

Figure 3. Demonstration of preprocessing (a) Original image with a horizontal thin line. (b) Binary image. (c) Eroded image which removes the background line. (d) Dilated and restored image.

image and B the structuring element which is chosen as disk type. The erosion of the binary image A by the structuring element B, denoted by $A\Theta B$, is defined as [13]

$$A\Theta B=\{z\,|\,(B)_z\subseteq A\} \tag{1}$$

where $(B)_z$ is the translation of B by the vector z, i.e., $(B)_z=\{c\,|\,c=a+z,a\in B\}$. The dilation of A by the structuring element B, denoted by $A\oplus B$, is defined as [13]

$$A\oplus B=\{z\,|\,(\hat{B})_z\ \mathrm{I}\ A\neq\varnothing\} \tag{2}$$

where is $\hat{B}$ the symmetric of B, that is,

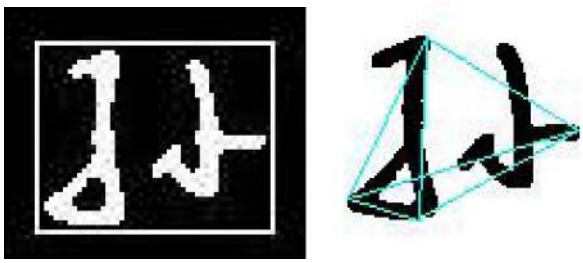
$$\hat{B}=\{w\,|\,w=-b,b\in B\}. \tag{3}$$

Figure 3 demonstrates the procedure of preprocessing. Given an original color image containing Chinese handwriting characters (Figure 3(a)), binary image can be obtained by directly applying binary operation as shown in Figure 3(b). Then erosion operation is further performed, through which the horizontal background line is removed as shown in Figure 3(c). The character is finally restored by the dilation operation, as shown in Figure 3(d). Since in most cases, strokes belonging to the same character are much closer than those belonging to different characters, single character can be extracted, which is further used in the feature extraction.

III. Extracting Features

Features are directly extracted from each single character. Since the stroke shapes and structures of Chinese characters are quite different from those of other languages such as English, where the handwriting characteristics are embedded, we propose to utilize the stroke shapes and structures for handwriting identification.

Through a number of experiments, we discover that the discriminatory handwriting characteristics lie in the two structures. They are the bounding rectangle and a special



(a) Bounding rectangle (b) TBLR quadrilateral

Figure 4. Two special structures of Chinese handwriting character. (a) Bounding rectangle. As we can see, it is the rectangle that exactly encloses the character. (b) TBLR quadrilateral. It is a quadrilateral that comprises four edge lines, as well as two diagonal lines, connecting four vertexes, i.e., Top-most, Bottom-most, Left-most, Right-most, thus has the name TBLR.

quadrilateral which we call TBLR quadrilateral, as shown in Figure 4(a) and Figure 4(b) respectively.

The following nine features are obtained from the bounding rectangle.

- $F1$: The ratio of the width to the height of the bounding rectangle, i.e.,

$$F1 = \frac{A_w}{A_h} \tag{4}$$

where A_w and A_h are the width and height of the bounding rectangle A respectively.

- $F2, F3$: The relative horizontal and vertical positions of the gravity center, i.e.,

$$F2 = \frac{\sum_{i=1}^{A_w} i \times P_x(i)}{\sum_{i=1}^{A_w} P_x(i)}, \tag{5}$$

$$F3 = \frac{\sum_{j=1}^{A_h} j \times P_y(j)}{\sum_{j=1}^{A_h} P_y(j)} \tag{6}$$

where $P_x(i)$ and $P_y(j)$ are the foreground pixel number in the i-th vertical and j-th horizontal line respectively.

- $F4, F5$: The relative horizontal and vertical gravity centers, i.e.,

$$F4 = \frac{F2}{A_w}, F5 = \frac{F3}{A_h}. \tag{7}$$

- $F6, F7$: The distance between the gravity center $G_1(x_1, y_1)$ and the geometric center $G_2(x_2, y_2)$, and the slope of the line connecting them, i.e.,

$$F6 = \sqrt{(y_2 - y_1)^2 + (x_2 - x_1)^2}$$
$$F7 = \frac{y_2 - y_1}{x_2 - x_1} \tag{8}$$

- $F8$: The ratio of the foreground pixel number to the area of the bounding rectangle, i.e.,

$$F8 = \frac{\sum_{i=1}^{A_w} \sum_{j=1}^{A_h} \times P(i, j)}{A_w \times A_h} \tag{9}$$

- $F9$: The stroke width property, i.e.,

$$F9 = \frac{\sum_{i=1}^{A_w} \sum_{j=1}^{A_h} \times P(i, j)}{\sum_{i=1}^{A_w} \sum_{j=1}^{A_h} \times P_t(i, j)} \tag{10}$$

where P_t is the binary pixel after refining the preprocessed image A. Given a structuring element $B = \{C, D\}$ consisting of two elements C and D, the refining operation keeps repeating the hit-or-miss operation, i.e., $A \mathrm{e} B = (A \Theta C) - (A \oplus \hat{D})$ until convergence, i.e., the change stops.

Similarly, from the TBLR quadrilateral, we can obtain the following seven features.

- $F10$: The ratio of the area of the top half part S_{up} to the area of the whole quadrilateral S , i.e.,

$$F10 = \frac{S_{up}}{S} \tag{11}$$

- $F11$: The ratio of the area of the left half part S_{left} to S , i.e.,

$$F11 = \frac{S_{left}}{S} \tag{12}$$

- $F12$: The cosine of the angle of the two diagonal lines, i.e.,

$$F12 = \cos(a, b) \tag{13}$$

where a and b are the direction vectors of the two diagonal lines respectively. The $F10, F11, F12$ measure the global spatial structure of the character.

- $F13$: The ratio of foreground pixel number P_{inner} within the TBLR quadrilateral to the total foreground pixel number P_{total} , i.e.,

$$F13 = \frac{P_{inner}}{P_{total}} \tag{14}$$

It measures the global degree of stroke aggregation.

- $F14$: The ratio of the P_{inner} to the area of the TBLR quadrilateral S_{TBLR} , i.e.,

$$F14 = \frac{P_{inner}}{S_{TBLR}} \tag{15}$$

- $F15$: The ratio of foreground pixel number of the left half part P_{left} within the TBLR quadrilateral to P_{total} , i.e.,

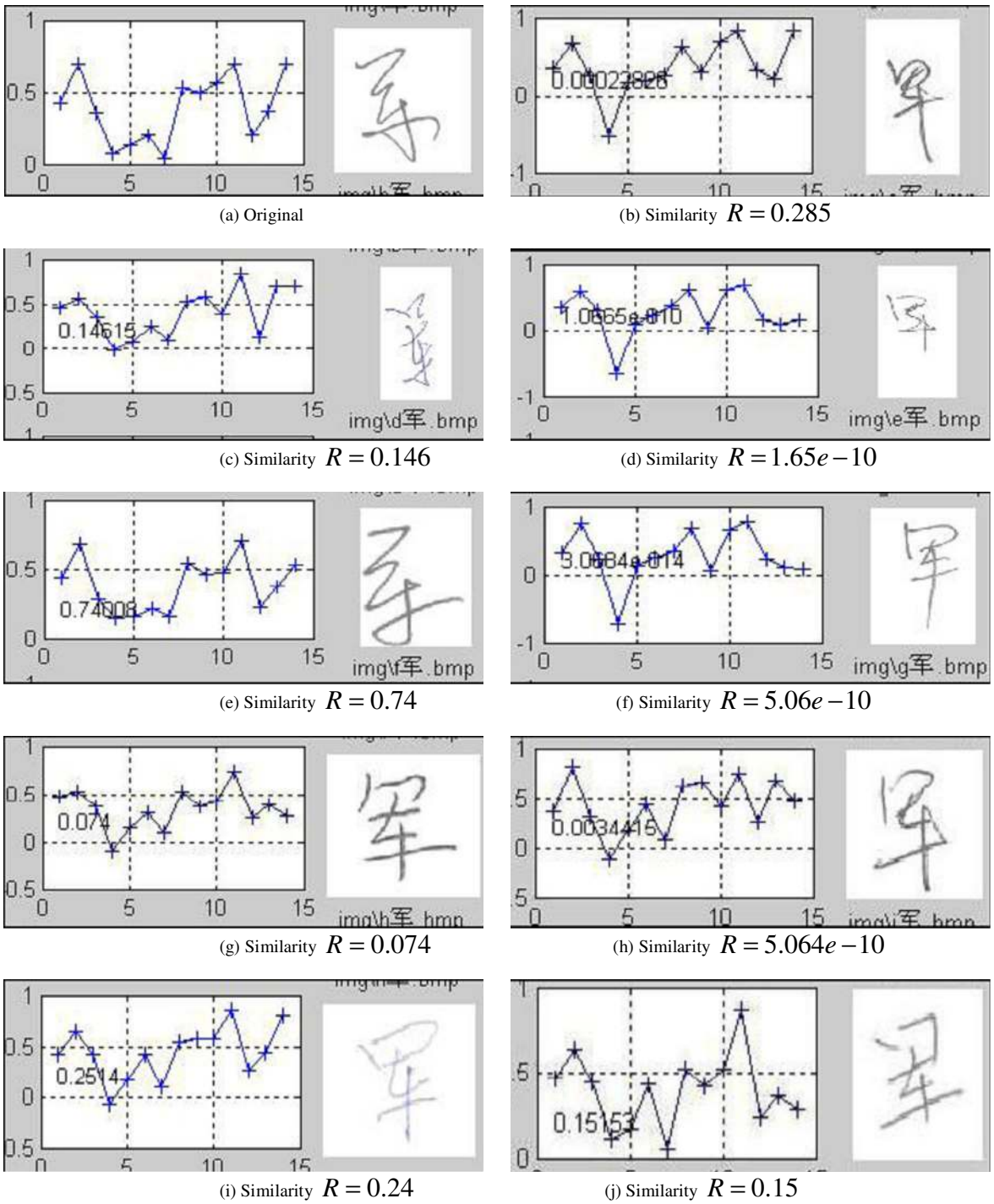
$$F15 = \frac{P_{left}}{P_{total}} \tag{16}$$

- $F16$: The ratio of foreground pixel number of the top half part P_{top} within the TBLR quadrilateral to P_{total} , i.e.,

$$F16 = \frac{P_{top}}{P_{total}} \tag{17}$$

Apart from the above sixteen features, we obtain another four features as follows.

- $F17$: The number of connected components. This feature measures the joined-up writing habit.
- $F18$: The number of hole within the character.
- $F19$: The number of stroke segments. It can be obtained by deleting all crossing point of a character, and the number is the total segment number.
- $F20$: The ratio of the longest stroke segment to the second longest stroke segment, where the stroke segments are obtained the same as that of $F19$.



(a) Original (b) Similarity $R = 0.285$

(c) Similarity $R = 0.146$ (d) Similarity $R = 1.65e-10$

(e) Similarity $R = 0.74$ (f) Similarity $R = 5.06e-10$

(g) Similarity $R = 0.074$ (h) Similarity $R = 5.064e-10$

(i) Similarity $R = 0.24$ (j) Similarity $R = 0.15$

Figure 5. Demonstration of the similarity. (a) The original (or saying, training sample) image. (b)-(j) are the testing samples appended with the similarity compared with (a). It can be seen that, the handwriting in (e) is the most similar to the original on the stroke shape and structure, and thus has the highest similarity.

These four features are used to adaptively adjust the similarity that is already evaluated from the first sixteen features as will be shown later.

IV. Matching

For each signer, thresholds of the twenty features are computed as the average feature values from the training characters of the same word by that signer, denoted as $a_k, k = 1,...,20$. Given a testing sample, twenty features are computed and compared with the corresponding thresholds of the same word. A set of

$r_k, k = 1,...,20$ are obtained first by

$$r_k = \begin{cases} \max\left[0, 1 - \frac{|Fk - a_k|}{a_k}\right] & \forall k = 1,...,16, \\ |Fk - a_k| \Big/ a_k & \forall k = 17,...,20 \end{cases} \quad (18)$$

Given a set of weights $c_k, k = 1,...,16$ which are computed from the training samples, the unadjusted similarity $\tilde{R}$ is computed as the weighted sum of the first sixteen features,

$$\tilde{R} = \sum_{i=1}^{16} r_k \times c_k. \quad (19)$$

Then the adjusted similarity R is adaptively obtained by

$$R = \begin{cases} \tilde{R} & \text{if } \exists k \in \{17,18,19,20\} \text{ s.t. } r_k < a_k \\ 0.9\tilde{R} & \text{otherwise} \end{cases} \quad (20)$$

Figure 5 illustrates the concept of similarity. It can be seen from the figures that, the similarity defined above has actually measured the degree of similarity of the stroke shape and structure. With a easily trained threshold, the assignment of each character to the correct signer can be obtained. For instance, the character in Figure 5(e) and the original shown in Figure 5(a) are considered being written by the same person if the similarity threshold is set at 0.7.

V. Experimental Results

In this section, we conduct experiments comparing the proposed method with the Fisher method and the method proposed by Bulacu and Schomaker [2], over two Chinese handwriting character databases. The experimental results have confirmed the effectiveness of our approach.

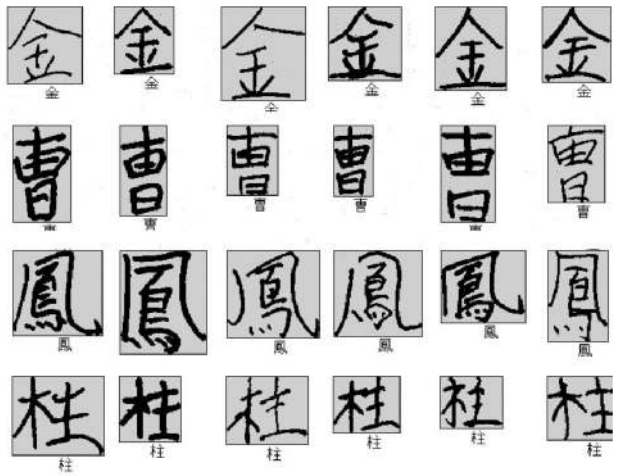


Figure 7. Examples of the HanjaDB1 database [19].

A. Databases

Two Chinese handwriting character databases are used.

Figure 6. Examples of the SYSU signature identification database.

The first database is the SYSU database which is generated and collected by ourselves as follows. 245 volunteers were asked to sign his (or her) name and one of the others' names twice. And a correction of 950 Chinese characters are obtained, which is named SYSU signature identification database. Figure 6 shows some examples of the SYSU database.

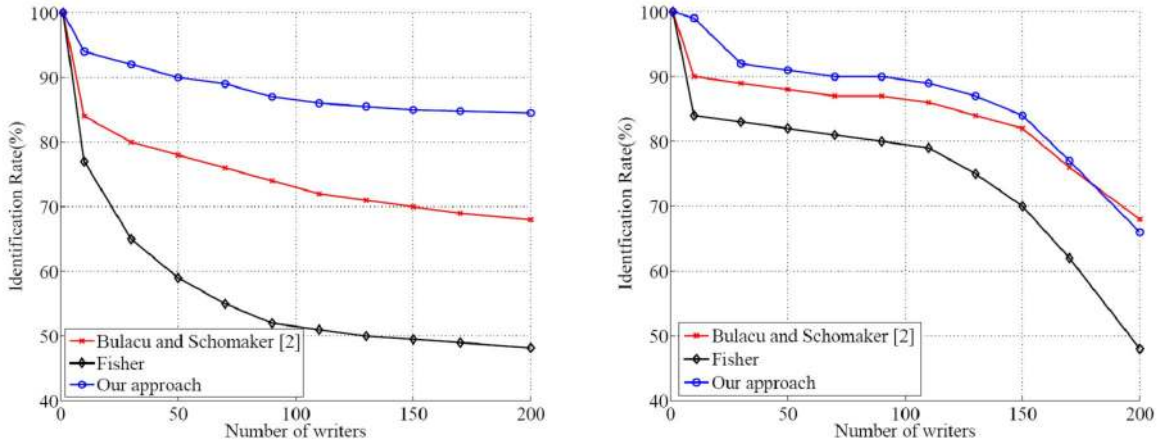
The second database is the HanjaDB1 database [19]. The images were written by more than 200 volunteers. The 800 most frequently used character classes in names of Korean, which covers 96.6% of usage, were collected. The subjects wrote Chinese characters on sheets containing 800 fields each of which was for one character. The sheets were scanned with flatbed scanner, and segmented into characters. The characters were filtered and sorted according to quality manually. Among 800 classes, 17 classes were discarded and 783 classes are remained. Figure 7 shows some examples of the HanjaDB1 database.

B. Comparison Results

Figure 8 plots the identification rate as a function of the number of writers by the three algorithms over two databases. In general, on both two databases, our approach has obtained the highest identification rate in all number of writers. Figure 9 shows the False Alarm Rate (FAR) and False Reject Rate (FRR) obtained by the three algorithms over the SYSU database. It can be seen that, the proposed method has again obtained both the lowest FAR and FRR on the SYSU database. The comparison results have validated the effectiveness of the proposed method.

VI. Conclusions

This paper presents a novel method for off-line Chinese handwriting identification. Two special structures, namely, the bounding rectangle and the TBLR quadrilateral, are explored to extract sixteen features.



(a) SYSU database (b) HanjaDB1 database

Figure 8. Comparing the identification rate as a function of the number of writers by the three algorithms over two databases.

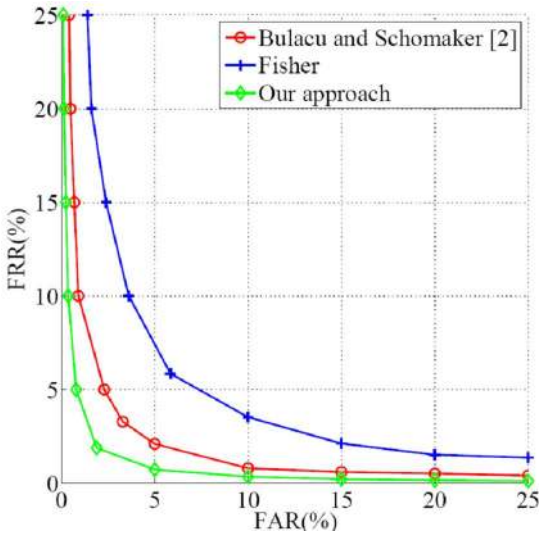


Figure 9. Comparing FAR and FRR obtained by the three algorithms on the SYSU database.

These sixteen features are used to compute the unadjusted similarity, which is further adaptively adjusted by another four commonly used features. The identification is directly performed on the adjusted similarity. Experiments on both SYSU and HanjaDB1 databases have been performed to compare the proposed method with two algorithms. Comparison results in terms of FAR, FRR and identification rate have confirmed the effectiveness of the propose approach.

ACKNOWLEDGMENT

This work was supported by the Science and Technology Program of Guangdong Province under Grant 2007B030603003. The authors would like to thank the Committees of ICIECS 2010 for recommending the paper to publish by the international journals. The authors would also like to thank KAIST for providing the HanjaDB1 Chinese handwriting database.

REFERENCES

[1] Z. Y. He, B. Fang, J. W. Du, Y. Y. Tang, and X. You, "A novel method for off-line handwriting-based writer identification," in Proc. of the 8th Int. Conf. on Document Analysis and Recognition, 2005.

[2] 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, April 2007.

[3] G. Zhu, Y. Zheng, D. Doermann, and S. Jaeger, "Signature detection and matching for document image retrieval," IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 11, pp. 2015– 2031, Nov. 2009.

[4] M. Panagopoulos, C. Papaodysseus, P. Rousopoulos, D. Dafi, and S. Tracy, "Automatic writer identification of ancient greek inscriptions," IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 8, pp. 1404–1414, Aug. 2009.

[5] T. Wang, Z. Chen, W. Li, X. Huang, P. Chen, and S. Zhu, "Relationship between personality and handwriting of Chinese characters using artificial neural network," in Proc. of the 1st Int. Conf. on Information Engineering and Computer Science, 2009.

[6] Z. Y. He and Y. Y. Tang, "Chinese handwriting-based writer identification by texture analysis," in Proc. of the 3rd Int. Conf. on Machine Learning and Cybemetics, 2004.

[7] Y. Zhu, T. Tan, and Y. Wang, "Biometric personal identification based on handwriting," in Proc. of the 15th Int. Conf. on Pattern Recognition, 2000.

[8] Q. Chen, Y. Yan, W. Deng, and F. Yuan, "Handwriting identification based on constructing texture," in Proc. of the 1st Int. Conf. on Intelligent Networks and Intelligent Systems, 2009.

[9] Z. Y. He, Y. Y. Tang, and X. You, "A contourlet-based method for writer identification," in Proc. of Int. Conf. on Systems, Man and Cybernetics, 2005.

[10] M. Bulacu, L. Schomaker, and L. Vuurpijl, "Writer identification using edge-based directional features," in Proc. of the 7th Int. Conf. on Document Analysis and Recognition, 2003.

[11] A. Schlapbach and H. Bunke, "Off-line handwriting identification using hmm based recognizers," in Proc. of the 17th Int. Conf. on Pattern Recognition, 2004.

[12] A. H.-R. Ko, P. R. Cavalin, R. Sabourin, and A. de Souza Britto Jr., "Leave-one-out-training and leave-one-out-testing hidden markov models for a handwritten numeral recognizer: The implications of a single classifier and multiple classifications," IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 12, pp. 2168–2178, Dec. 2009.
[13] R. C. Gonzalez and R. E. Woods, Digital Image Processing, 2nd ed. Prentice Hall, 2002.
[14] J. Tan, J.-H. Lai, C.-D. Wang and M.-S. Feng, "Off-line Chinese Handwriting Identification Based on Stroke Shape and Structure," in Proc. Of the 2nd Int. Conf. on Inf. Engineering and Computer Science, 2010.
[15] E. N. Zois and V. Anastassopoulos, "Morphological waveform coding for writer identification," Pattern Recognition, vol. 33, pp.385-398, 2000.
[16] W. Y. Leng and S. M. Shamsuddin, "Writer identification for Chinese handwriting," Int. J. Advance. Soft Comput. Appl., vol. 2, no. 2, July, 2010.
[17] F. H. Cheng, "Multi-stroke relaxation matching method for handwritten Chinese character recognition," Pattern Recognition. Vol. 31, no. 4, pp.401-410, 1998.
[18] X. Li and X. Ding, "Writer Identification of Chinese Handwriting Using Grid Microstructure Feature," ICB, 2009.
[19] http://ai.kaist.ac.kr/Resource/dbase/KAIST-DB.htm#Hanja

Jun Tan received the B.Sc. and M.Sc. degrees in computational mathematics from Sun Yat-sen University, Guangzhou, P. R. China, in 1995 and 2001, respectively. He started the pursuit of the Ph.D. degree with the School of Information Science and Technology of Sun Yat-Sen University in September 2009.
He did teaching and research in the department of scientific computation and computer application, Sun Yat-Sen University, Guangzhou, P. R. China, from 2002.
Mr. Tan is currently working on developing statistical pattern recognition methods for automatic writer identification and for handwritten historical document retrieval. His scientific interests include computer vision, statistical pattern recognition, biometrics, and document analysis and recognition.

Jian-Huang Lai received his M.Sc. degree in applied mathematics in 1989 and his Ph.D. in mathematics in 1999 from SUN YAT-SEN University, P. R. China.
He joined Sun Yat-sen University in 1989 as an Assistant Professor, where currently, he is a Professor with the Department of Automation of School of Information Science and Technology and vice dean of School of Information Science and Technology.
Prof. Lai had successfully organized the International Conference on Advances in Biometric Personal Authentication'2004, which was also the Fifth Chinese Conference on Biometric Recognition (Sinobiometrics'04), Guangzhou, in December 2004. He has taken charge of more than five research projects, including NSF-Guangdong (No.U0835005), NSFC (NO.60144001, 60373082, 60675016), the Key (Keygrant) Project of Chinese Ministry of Education (No.105134), and NSF of Guangdong, China (No.021766, 06023194). He has published over 80 scientific papers in the international journals and conferences on image processing and pattern recognition. His current research interests are in the areas of digital image processing, pattern recognition, multimedia communication, wavelet and its applications.
Prof. Lai serves as a standing member of the Image and Graphics Association of China and also serves as a standing director of the Image and Graphics Association of Guangdong.

Chang-Dong Wang received the B.S. degree in applied mathematics in 2008 and M.Sc. degree in computer science in 2010 from Sun Yat-sen University, Guangzhou, P. R. China. He started the pursuit of the Ph.D. degree with Sun Yat-sen University in September 2010.
He has published several scientific papers in the international journals and conferences on image processing and pattern recognition. He was selected for the IEEE TCII Student Travel Award and gave a 20 minutes' oral presentation in the 10th IEEE International Conference on Data Mining, December 14-17, 2010, Sydney, Australia. His ICDM paper titled "A Conscience On-line Learning Approach for Kernel-Based Clustering" has been selected as one of the four best research papers and invited to be extended for publication in the international journal Knowledge and Information Systems.
Mr. Wang's current research interests include machine learning, pattern recognition and computer vision, especially focusing on data clustering and its applications in computer vision. He is a student member of IEEE.

