Chinese Handwritten writer identification based on Structure Features and extreme learning machine

Jun Tan
Guangdong Province Key Laboratory of
Computational Science
Sun Yat-Sen University
GuangZhou,PR China 510275
Email: mcstj@mail.sysu.edu.cn

JianHuang Lai
School of Information and
Science and Technology
Sun Yat-Sen University
GuangZhou,PR China 510275
Email: stsljh@mail.sysu.edu.cn

Wei-Shi Zheng
School of Information and
Science and Technology
Sun Yat-Sen University
GuangZhou,PR China 510275
Email: wszheng@ieee.org

Abstract—In this paper,we propose a new approach for writer identification of Chinese handwritten.In our method, we deal with writer identification of Chinese handwritten using Chinese character structure features(CSF) and extreme learning machine(ELM).To extract the features embedded in Chinese handwriting characters, special structures have been explored according to the trait of Chinese handwriting characters,where 20 features are extracted from the structures, these features constitute patterns of writer handwriting. We also combine structure features with extreme learning machine (ELM) as a new scheme for writer recognition, ELM is single hidden layer feed forward networks (SLFN), which randomly chooses the input weights and analytically determines the output weights of SLFN. This algorithm learns much faster than traditional popular learning algorithms. Experimental results demonstrate CSF/ELM method can achieve better performance than other traditional schemes for writer identification.

I. INTRODUCTION

As one of the most important methods in the biometric individual identification, writer identification has been widely used in the fields of bank check, forensic, historic document analysis, archaeology, identifying personality [1], many approaches have been developed [2], [3]. According to the different input methods, writer identification is commonly classified into on-line and off-line.

Compared with its on-line counterpart, off-line writer identification is a rather challenging problem. Chinese characters are ideo graphic in nature[4], 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 [5], in Chinese characters, the complication structures are mostly affected by multi stokes of each character. Additionally, as shown in Fig.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 [2]. The approaches proposed for English handwriting writer identification is no longer suitable for the case of Chinese handwritings [6], [7]. In this paper, we propose Chinese structure feature(CSF) as algorithm of feature extraction and combine CSF with extreme learning machine (ELM) as a new scheme for writer identification.

(a) Sample of Chinese handwriting   (b) Sample of English handwriting

Fig. 1: Samples of the Chinese and English handwritings

A. Related Work

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

Given a free style handwritten document, a preprocessing is often required. Segmentation is an indispensable step in preprocessing. Some methods have been proposed to segment characters recently [8], we proposed a method for Chinese character segmentation based on nonlinear clustering[9].

Handwriting features except CSF feature[10] others such as texture, edge, contour and character shape have been widely studied recently. Several researchers [11], [12] proposed to take the handwriting as an image containing special texture, and therefore regarded writer identification as the texture identification. Among them, Zhu [11] adopted 2-D Gabor filtering to extract the texture features, while Chen et al. [12] used the Fourier transform. Xu and Ding[13] proposed a histogram-based feature to identify writer, called grid microstructure feature which is extracted from the edge image of the scanned images.

In [10], we propose a method for extracting Chinese structure features(CSF). Despite good performance, one serious drawback is that, it only compare one by one sample using algorithm of Similarity Matching, and it cannot classify multi-samples to different writer class. Several classifier methods have been developed to overcome the problem.

Once discriminant features have been extracted, they are submitted to a classifier whose task is to identify writer that they represent. The widely used classifiers at least include Weighted Euclidean Distance(WED) classifier [2], [11], [12], Bayesian model,BP neural networks [3], likelihood ranking [14], SVM[15]. For matching singleton non-sequential features such as texture, edge and contour, Weighted Euclidean Distance (WED) has been shown to be effective by the experiments. In [3], both Bayesian classifiers and neural networks were used as the classifiers. Imdad [15] use Steered Hermite Features to identify writer from a written document, and the algorithm takes Support Vector Machine for training and testing.

The traditional algorithms for this issue such as back-propagation (BP) need many iterative steps to calculate the optimal values of the input weights and the output weights, so their speeds are very slow in general. ELM [16] is an efficient and practical learning mechanism for the single-hidden-layer feed-forward neural networks. ELM[17] can learn the input weights and the output weights by directly calculating the Moore$-$Penrose generalized inverse matrix of the hidden layer output matrix of the neural network instead of using the iterative steps. So, it is necessary to perform efficient features extraction on the one hand, and to take steps to reduce the training/testing time on the other hand[18]. ELM is an efficient algorithm which tends to reach the smallest training error, obtain smallest norm of weights, produce best generalization performance, and runs faster than the iterative algorithms[19].

The rest of the paper is organized as follows. In Sect. II, we first briefly review algorithms of CSF feature extraction, ELM is briefly explained in Sect.III. Our proposed scheme in Sect.IV. We analyze the experimental results in Sect.V. Finally, the conclusion is given in Sect. VI.

II. CHINESE STRUCTURE FEATURES(CSF)

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[10]. They are the bounding rectangle and a special quadrilateral which we call TBLR quadrilateral, as shown in Fig.2(a) and Fig.2(b) respectively.

The following nine Chinese Structure features(CSF) are obtained from the bounding rectangle as shown in Table I. $F1$: The ratio of the width to the height of the bounding rectangle A; $F2, F3$: The relative horizontal and vertical positions of the gravity center; $F4, F5$: The relative horizontal and vertical gravity centers; $F6, F7$: 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; $F8$: The ratio of the foreground pixel number to the area of the bounding rectangle; $F9$: The stroke width property,where P_t is the binary pixel

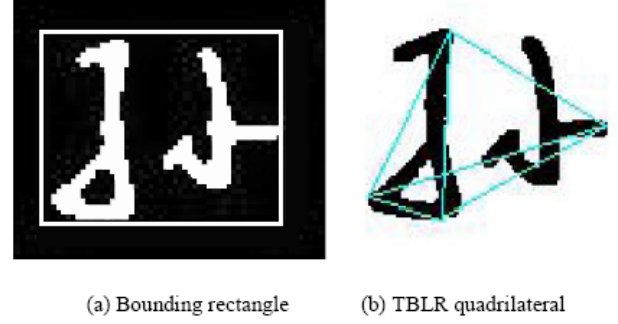(a) Bounding rectangle   (b) TBLR quadrilateral

Fig. 2: Two special structures of Chinese handwriting character. (a) Bounding rectangle. (b) TBLR quadrilateral.

TABLE I: CSF feature from the bounding rectangle

ith	Eqs.	Comments
1	A_w/A_h	A_w andA_h are the width and height of A.
2	$\frac{\sum_{i=1}^{A_w} i \times P_x(i)}{\sum_{i=1}^{A_w} P_x(i)}$	Foreground pixel number $i-$th vertical$P_x(i)$
3	$\frac{\sum_{j=1}^{A_h} j \times P_y(j)}{\sum_{j=1}^{A_h} P_y(j)}$	Foreground pixel number $j-$th horizontal$P_y(i)$
4	$F2/A_w$	$F2$ is 2th CSF feature
5	$F3/A_h$	$F3$ is 3th CSF feature
6	$\|G1 - G2\|$	Gravity center $G_1(x_1, y_1)$
7	$(y_2 - y_1)/(x_2 - x1)$	Geometric center$G_2(x_2, y_2)$
8	$\frac{\sum_{i=1}^{A_w} \sum_{j=1}^{A_h} \times P(i,j)}{A_w \times A_h}$	Foreground pixel number $P(i,j)$
9	$\frac{\sum_{i=1}^{A_w} \sum_{j=1}^{A_h} \times P(i,j)}{\sum_{i=1}^{A_w} \sum_{j=1}^{A_h} \times P_t(i,j)}$	Binary pixel after refining$P_t(i,j)$

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,

$$A \circledast B = (A\Theta C) - (A \oplus \hat{D}) \qquad (1)$$

until convergence, i.e., the change stops.

Similarly, from the TBLR quadrilateral, we can obtain the following seven CSF features as shown in Table II. $F10$: The ratio of the area of the top half part $S_{up}$to the area of the whole quadrilateral S ; $F11$: The ratio of the area of the left half part S_{left} to S ; $F12$: The cosine of the angle of the two diagonal lines,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} . It

TABLE II: CSF features from the TBLR quadrilateral

ith	Eqs.	Comments
10	S_{up}/S	S_{up} is the area of the top half part.
11	S_{left}/S	S_{left} is the area of left half part.
12	$\cos(a,b)$	a and b are the direction vectors of diagonal
13	P_{inner}/P_{total}	Foreground pixel number P_{inner}
14	P_{inner}/S_{TBLR}	Area of the $TBLR$ quadrilateral S_{TBLR}
15	P_{left}/P_{total}	Foreground pixel number of left half part P_{left}
16	P_{top}/P_{total}	Foreground pixel number of top half part P_{top}

measures the global degree of stroke aggregation. $F14$: The ratio of the P_{inner} to the area of the $TBLR$ quadrilateral S_{TBLR}; $F15$: The ratio of foreground pixel number of the left half part P_{left} within the $TBLR$ quadrilateral to P_{total}; $F16$: The ratio of foreground pixel number of the top half part P_{top} within the $TBLR$ quadrilateral to P_{total}.

Apart from the above sixteen features, we obtain another four CSF 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 $F19$.

III. EXTREME LEARNING MACHINE(ELM)

For N arbitrary distinct samples (X_i, T_i), where $X_i = [x_{i1}, x_{i2}, \ldots, x_{in}]^T \in R^n$ and $T_i = [t_{i1}, ti2, \ldots, t_{im}]^T \in R^m$, standard SLFN with $\hat{N}$ hidden neurons and activation function g(x) are mathematically modeled as follow:

$$\sum_{i=1}^{\hat{N}} \beta_i g(W_i \cdot X_j + b_i) = O_j, j = 1, 2, ..., N \qquad (2)$$

where $W_i = [w_{i1}, w_{i2}, ..., w_{in}]^T$ is the weight vector connecting the ith hidden neuron and the input neurons, $\beta_i = [\beta_{i1}, \beta_{i2}, ..., \beta_{im}]^T$ is the weight vector connecting the ith hidden neuron and the output neurons, and b_i is the threshold of the ith hidden neuron. The numbers of input and output neurons are represented using n and m respectively. $W_i \cdot X_j$ denotes the inner product of W_i and X_j. The output neurons are chosen linear in this experiment.

The architecture of ELM classifier is shown in Fig.3. In the training procedure, the SLFN with $\hat{N}$ hidden neurons with activation function $g(x)$ can approximate these N samples with zero error means that $\sum_{i=1}^{\hat{N}} \|o_j - t_j\| = 0$ i.e., there exist β_i, W_i and b_i such that

$$\sum_{i=1}^{\hat{N}} \beta_i g(W_i \cdot X_j + b_i) = t_j, j = 1, 2, ..., N \qquad (3)$$

The above N equations can be written compactly as:

$$\mathbf{H}\beta = \mathbf{T} \qquad (4)$$

where

$$\mathbf{H}(W_1, ..., W_{\hat{N}}, b1, ..., b_{\hat{N}}, X_1, ..., X_N)$$
$$= \begin{bmatrix} g(W_1 \cdot X_1 + b_1) & ... & g(W_{\hat{N}} \cdot X_1 + b_{\hat{N}}) \\ \vdots & \vdots & \vdots \\ g(W_1 \cdot X_N + b_1) & ... & g(W_{\hat{N}} \cdot X_N + b_{\hat{N}}) \end{bmatrix}_{N \times \hat{N}} \qquad (5)$$

$$\beta = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_{\hat{N}}^T \end{bmatrix}_{\hat{N} \times m} \quad and \quad \mathbf{T} = \begin{bmatrix} t_1^T \\ \vdots \\ t_N^T \end{bmatrix}_{N \times m} \qquad (6)$$
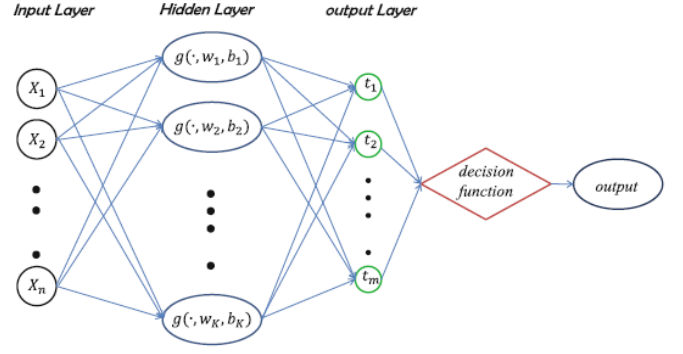

Fig. 3: The structure of Extreme Learning Machine classifier.

$\mathbf{H}$ is called the hidden layer output matrix of the neural network; the ith column of $\mathbf{H}$ is the ith hidden neuron's output vector with respect to inputs $X_1, X_2, \ldots, X_N$[16].

If the number of hidden neurons is equal to the number of distinct training samples,i.e.$\hat{N} = N$, matrix $\mathbf{H}$ is square and invertible, and SLFN can approximate these training samples with zero error. However, in most cases the number of hidden neurons is much less than the number of distinct training samples, $\hat{N} \ll N$, so $\mathbf{H}$ is a non square matrix and there may not exist $W_i, b_i, \beta_i(i = 1, \ldots, \hat{N})$ such that $\mathbf{H}\beta = T$. Thus, one may need to find specific $\hat{W}_i, \hat{b}_i, \hat{\beta}_i(i = 1, \ldots, N)$ such that

$$\|\mathbf{H}(\hat{W}_1, \ldots, \hat{W}_{\hat{N}}, \hat{b}_1, \ldots, \hat{b}_{\hat{N}})\hat{\beta} - \mathbf{T}\|$$
$$= \min_{W_i, b_i, \beta_i} \|\mathbf{H}(W_1, \ldots, W_{\hat{N}}, b_1, \ldots, b_{\hat{N}})\beta - \mathbf{T}\| \quad (7)$$

the smallest norm least squares solution of the above linear system is:

$$\hat{\beta} = \mathbf{H}^\dagger \mathbf{T} \qquad (8)$$

Algorithm ELM: Given a training set $\aleph = (\mathbf{x}_i, \mathbf{t}_i)|x_i \in R^n, t_i \in R^m, i = 1, \ldots, N$, activation function $g(x)$ and hidden neuron number $\hat{N}$,

Step 1: Assign arbitrary input weight $\mathbf{w}_i$ and bias b_i, $i = 1, \ldots, \hat{N}$

Step 2: Calculate the hidden layer output matrix $\mathbf{H}$.
Step 3: Calculate the output weight $\hat{\beta}$ by Eqs.(8), where $\mathbf{H}$ and $\mathbf{T}$ are defined as Eqs. (5) and (6).

IV. OUR SCHEME

Some of the results in this paper were first presented in [10]. In this paper, we present more technique details effectiveness of CSF/ELM approach. Fig.4 demonstrate the flowchart of the proposed approach. There are three main steps for Chinese handwritten writer identification. The first step is handwritten image preprocessing, which removes noises and normalizes the images into the same size. The second step is feature extraction, which finds effective representation of the difference of writers in handwritten. Instead of using complex feature extraction methods, we propose Chinese structure features(CSF) for feature extraction. The last step is to apply ELM learning method to classify different writers.
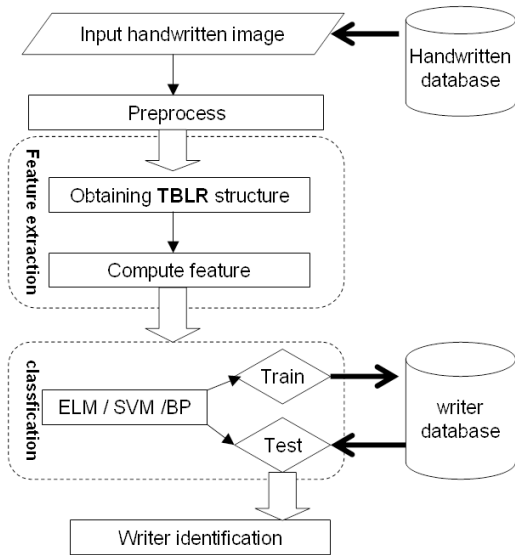
Fig. 4: The flowchart of the proposed approach.

For example, the entire process of CSF/ELM-based handwritten writer recognition is as follows:

- Step 1: The appropriate training strategy based on the selected training set, we randomly selected from a handwritten database as part of the training set $TrainSet = Si, i = 1, ..., N$, where N is the total of training samples, and the remaining samples as the test set;
- Step 2: Image pre-processing for training set, through the noise removal and standardization;
- Step 3: CSF feature extraction method to extract the optimal recognition feature vector, 20 features are extracted from structures of Bounding Box and TBLR quadrilateral;
- Step 4: ELM train phase, using **Algorithm ELM** and Eqs. (5)(6)(9), set the input weight parameters arbitrary $\mathbf{w}_i$ and bias $b_i, i = 1, \ldots, \hat{N}$ randomly, get the hidden layer output matrix $\mathbf{H}$ and the output weight β, the model of training has been trained, the training process is complete;
- Step 5: ELM test phase. Testing the model parameters obtained from the training model, and then we can obtain the actual output through the test image by the Eqs. (3), to identify the writer.

V. EXPERIMENTS AND ANALYSIS

A. Handwritten Database

To test the performance of the proposed method in the writer identification, we do some experiments over 2 Chinese handwritten databases: SYSU [10] and KAIST Hanja1 [20]. Among them, SYSU database which was 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. The KAIST Hanja1 database contains 783 frequently used Chinese characters, where each character consisting of 200 samples written by 200 writers respectively. Fig.5 and Fig.6 shows some samples of the databases.
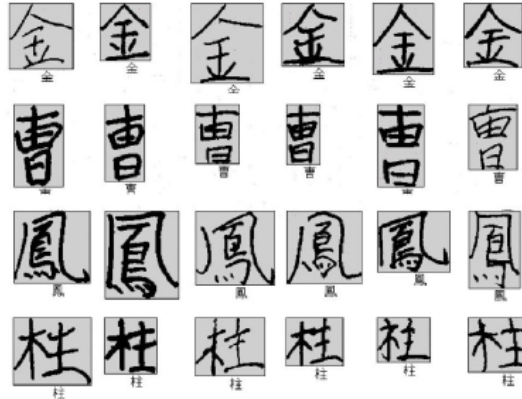

Fig. 5: Examples of the HanjaDB1 database.

Fig. 6: Examples of the SYSU database.

B. Comparative results

In experiment, because the features may have large differences in value, in order to avoid large values of features to submerge the contributions of the small value of features, all samples were normalized between 0 and 1 before sending to the learning algorithms as input.

We compare the proposed method with three well-known methods including the methods in[11], [12], [13]. Each of the compared methods is well-adjusted/trained to generate the best results. Both the recognition accuracy of writer identification and average time cost are reported and compared.

Furthermore, learning method is also an important problem. We used learning methods(ELM,BP[3],SVM[15]) for testing, and the average training time and the average test time is calculated. The cost time of the experiments is shown in table III. From the table, we can see that SVM[15] and BP[3] training times are relative much more than the ELM training time,the average total time of ELM, SVM, BP are 1.3001, 31.395, 34.801 seconds respectively. Therefore, the method of ELM has the highest speed.

TABLE III: Time cost (seconds) of different learning methods

Method	Average training time	Average test time	Average total time
ELM	1.412	0.0312	1.3001
SVM[15]	32.05	0.063	31.395
BP[3]	36.37	0.382	34.801

TABLE IV: Recognition accuracies of different methods

Method	Top1 Hanja	Sysu	Top5 Hanja	Sysu	Top10 Hanja	Sysu	Top20 Hanja	Sysu
Gabor[11]	54.3	53.2	61.3	64.2	67.2	68.8	72.2	73.9
Fourier[12]	42.7	52.3	54.5	61.7	64.8	70.2	72.3	74.5
GMSF[13]	74.5	75.4	82.7	83.2	85.6	88.4	87.2	91.4
Our	81.3	82.6	86.3	87.5	91.2	92.2	95.4	98.5

Finally, we compare our scheme with approaches [11], [12], [13]in the Chinese handwritten database Hanja and SYSU. These approaches use different features and classifiers. method[11] using Gabor feature and WED classifier, GMSF method[13] using GMSF feature and Weighted Chi-square, and Fourier feature and Mathematical expectations classifier are for method[12]. In the comparison, Table IV gives the Top-1, Top-5, Top-10 and Top-20 recognition accuracies of the four methods. We show the recognition accuracies of the algorithms on the handwritten database in Table IV. From the table, we can see the accuracies of our method are higher than the others. The lowest accuracy of method[12] is 42.7%, and our method has the similar accuracy of method[13]. It is obvious that our method is more effective for identifying writer in Chinese handwriting.

VI. CONCLUSION

In this paper, we list some results of the literatures. Gabor and wavelet features[11] used in the traditional methods of Chinese writer identification are affected greatly by the normalization and the arrangement of characters in texture blocks. Differently, CSF feature uses original handwriting images and tries to find out the writing structure of the writer in local regions. The ELM is a classifier used to train a single hidden layer neural network. From the experimental results. We can see that, Table IV includes the performance of our method and some other methods for Chinese writer identification. The recognition accuracy of our method using the CSF/ELM seems better than the existing methods for Chinese writer identification. compared with traditional learning algorithm, ELM has faster speed, better generalization performances. The effectiveness of CSF/ELM for Chinese writer identification is proved by the experiments.

It is expectable that our approach can be used for multilingual handwriting including western handwritings and arabic number.

ACKNOWLEDGMENT

REFERENCES

[1] S. Impedovo and G. Pirlo. Verification of handwritten signatures: an overview. In *Image Analysis and Processing, 2007. ICIAP 2007. 14th International Conference on*, pages 191–196. IEEE, 2007.

[2] Z. He, Y.Y. Tang, and X. You. A contourlet-based method for writer identification. In *Systems, Man and Cybernetics, 2005 IEEE International Conference on*, volume 1, pages 364–368. IEEE, 2005.

[3] M. Bulacu, L. Schomaker, and L. Vuurpijl. Writer identification using edge-based directional features. *writer*, 1:1, 2003.

[4] L. Shan. Passport to chinese: 100 most commonly used chinese characters, book 1, 1995.

[5] F.H. Cheng. Multi-stroke relaxation matching method for handwritten chinese character recognition. *Pattern recognition*, 31(4):401–410, 1998.

[6] A. Schlapbach and H. Bunke. Off-line handwriting identification using hmm based recognizers. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 2, pages 654–658. IEEE, 2004.

[7] L. Schomaker and L. Vuurpijl. Forensic writer identification: A benchmark data set and a comparison of two systems [internal report for the {Netherlands Forensic Institute}]. 2000.

[8] L.Y. Tseng and R.C. Chen. Segmenting handwritten chinese characters based on heuristic merging of stroke bounding boxes and dynamic programming1. *Pattern Recognition Letters*, 19(10):963–973, 1998.

[9] J. Tan, J.H. Lai, and C.D.; Wang. A new handwritten character segmentation method based on nonlinear clustering. *NEUROCOMPUTING*, 89(8):213–219, 2012.

[10] J. Tan, J.H. Lai, C.D. Wang, and M.S. Feng. A stroke shape and structure based approach for off-line chinese handwriting identification. *International Journal of Intelligent Systems and Applications (IJISA)*, 3(2):1, 2011.

[11] Y. Zhu, T. Tan, and Y. Wang. Biometric personal identification based on handwriting. In *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, volume 2, pages 797–800. IEEE, 2000.

[12] Q. Chen, Y. Yan, W. Deng, and F. Yuan. Handwriting identification based on constructing texture. In *Intelligent Networks and Intelligent Systems, 2008. ICINIS'08. First International Conference on*, pages 523–526. IEEE, 2008.

[13] Lu Xu, Xiaoqing Ding, Liangrui Peng, and Xin Li. An improved method based on weighted grid micro-structure feature for text-independent writer recognition. In *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, pages 638–642. IEEE, 2011.

[14] G. Zhu, Y. Zheng, D. Doermann, and S. Jaeger. Signature detection and matching for document image retrieval. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(11):2015–2031, 2009.

[15] A. Imdad, S. Bres, V. Eglin, H. Emptoz, and C. Rivero-Moreno. Writer identification using steered hermite features and svm. In *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, volume 2, pages 839–843. IEEE, 2007.

[16] G.B. Huang, Q.Y. Zhu, and C.K. Siew. Extreme learning machine: theory and applications. *Neurocomputing*, 70(1):489–501, 2006.

[17] N.Y. Liang, G.B. Huang, P. Saratchandran, and N. Sundararajan. A fast and accurate online sequential learning algorithm for feedforward networks. *Neural Networks, IEEE Transactions on*, 17(6):1411–1423, 2006.

[18] X.Z. Wang and C.R. Dong. Improving generalization of fuzzy if–then rules by maximizing fuzzy entropy. *Fuzzy Systems, IEEE Transactions on*, 17(3):556–567, 2009.

[19] G.B. Huang, D.H. Wang, and Y. Lan. Extreme learning machines: a survey. *International Journal of Machine Learning and Cybernetics*, 2(2):107–122, 2011.

[20] I.J. Kim and J.H. Kim. Statistical character structure modeling and its application to handwritten chinese character recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(11):1422–1436, 2003.