Personal Handwriting Identification Based on PCA

Long Zuo, Yunhong Wang, Tieniu Tan*
National Laboratory of Pattern Recognition (NLPR)
Institute of Automation, Chinese Academy of Sciences

ABSTRACT

In this paper, a novel algorithm is presented for writer identification from handwritings. Principal Component Analysis is applied to the gray-scale handwriting images to find a set of individual words which best characterize a person's handwriting style and have maximal difference from other people style. During identification, we only need to utilize a set of individual characteristic words for comparison, instead of comparing the whole handwriting text to identify the writers. So not only is a very high average identification performance of 97.5% obtained, but also a very fast identification speed is achieved in our method.

In the experiment, 400 pages of handwriting texts, containing almost 16000 Chinese words written by 40 different writers are used to validate the performance of the method.

Keywords: biometrics, writer identification, principal component analysis (PCA), characteristic words

1. Introduction

Biometrics has been an active research area for a long time aiming at automatic identity recognition based on individual physiological or behavioral characteristics. Security is an important issue in an increasingly networked environment, and accurate identification of people could deter crime and fraud, improve the living quality and save critical resources. In order to ensure convenience and safety in economic activities, it is necessary to perform accurate and rapid personal identification.

Personal identification based on handwriting is a kind of behavioral biometric identification approach. Each person has his individual writing style and the handwritings are easy to obtain. For this reason, much research has touched on this field, but most of them rested on signature verification, which has the disadvantage in that the identification content is fixed and limited, making it prone to forgery.

Fundamentally different from the previous methods, the algorithm presented in this paper is based on the fact that each writer's handwriting style can sufficiently be characterized by a few words which are called characteristic words in this paper. We just use these words to recognize the writer's identity. The characteristic words are acquired by comparing the average distance between the feature vectors projected onto the PCA subspaces.

The rest of this paper is organized as follows. In Section 2, an overview of related existing methods is given. Section 3 introduces our new algorithm and Section 4 presents the experimental results. Section 5 summarizes this paper.

2. Overview of related work

There have been a few attempts in writer identification in the past few years. The texture analysis approach to writer identification was proposed by Said, Tan and Baker[2]. In this method, they took the handwriting as an image containing some special texture, and applied the well-established 2-D Gabor filtering technique to extract features of such textures. This was a content independent method. Shrihari and Cha [7] extracted twelve shape features from the handwriting text lines to represent personal handwriting style. The features mainly contained visible characteristics of the handwriting, such as width, slant and height of the main writing zones. Some other papers also adopted multiple features integration to writer identification [6][7]. In reference [5] and [6], two writer verification systems were proposed, judging an individual identity based on a short sentence and each word in the sentence is used to tackle the individual verification problem. So the proposed method makes it more difficult for the forgery to simulate the authentic

* { E_mails:{lzuo,wangyh, tnt}@nlpr.ia.ac.cn; phone 86-10-62647441; fax 86-10-62551993; http://www.sinobiometrics.com/; National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, Beijing, China.

writer than the previous methods using only signature.

In general, most of the above methods assumed that handwriting images are binary images. However, along with the development of human-machine interaction equipments, especially with the development of the high sensitive pen and many delicate data collection systems [8], more handwriting style information can be easily collected while writing. The writing pressure information is an important feature, which can be represented as gray-scale images for personal authentication. Moreover, the above methods typically compared the whole handwriting text to determine whether the two texts are written by the same person. However, from our daily experience, writers may be distinguished based only on a very small number of characteristic words. In the following, we explore this fact for efficient characteristic words based writer identification.

3. The new algorithm

3.1 Overview of our method

Our idea originates from the well-known PCA based face recognition method. But different from face recognition, here the same word from each writer's text is trained to form a PCA subspace. So we obtained a number of subspaces, and each subspace corresponds to a word in the handwriting text. The word features are represented in a reduced dimensionality, and we design an algorithm to extract each writer's characteristic words which can best characterize the different writing style from one writer to another. Then correspondence between the writer and his characteristic words is established. For the unknown writer's handwriting, features are compared with those of the known writers through characteristic words instead of comparing the whole handwriting texts.

3.2 Data collection and processing

We establish our handwriting database by using a Wacom PL-400 digitizing tablet with the UP 811E pressure sensitive pen recording pen location and pressure at a sampling rate of 20 samples per second. Hence, we obtain the handwriting gray-scale images containing pressure information. Figure 1 shows some examples of handwriting texts.

(a) (b)

Figure 1. Some examples of handwriting texts

In the experiment, we request the participants to write a content-fixed text containing 40 Chinese words, and each participant wrote the text 10 times. Consequently, 400 handwriting texts were recorded in our database containing a total of 16000 Chinese words.In the preprocessing step, the main task is to apply linear normalization algorithm to reduce the effects of size variations in the handwriting. One point to be noticed here is that we do not carry out skewness correction, because skewness is also a feature in writer's handwriting style.

3.3 Determinant of characteristic words

The normalized handwriting word images are taken as the training set. The whole training algorithm and computational considerations are as follows.

First, an input word image of size w×h pixels is represented as an n-dimensional (n= w×h) vector $x_{i,j,k}$ by concatenating the rows of the image, where $x_{i,j,k}$ means the k-th word written by the i-th person in the j-th writing.For the PCA, the sample covariance matrix of input word images can be computed according to the follow expression:

$$\Sigma_k = \frac{1}{I \times J} \sum_{i=1}^{I} \sum_{j=1}^{J} (x_{i,j,k} - \mu_k)(x_{i,j,k} - \mu_k)^T \tag{1}$$

where $\mu_k = \frac{1}{M} \sum_{i=1}^{I} \sum_{j=1}^{J} x_{i,j,k}$ represents the mean of each word.

Next, we determine the eigenvalues and eigenvectors of the covariance matrix. If the rank of the matrix Σ_k is N_k, we can compute N_k nonzero eigenvalues $\lambda_{k,1} > \lambda_{k,2} > \cdots > \lambda_{k,N_k}$ and the associated eigenvectors $e_{k,1}, e_{k,2} \cdots$

e_{k,N_k} . Generally speaking, we only need chose the first few largest eigenvalues and their corresponding eigenvectors using a threshold value α:

$$\frac{\sum_{s=1}^{N_l}\lambda_{k,N_s}}{\sum_{s=1}^{N_k}\lambda_{k,N_s}} > \alpha \tag{2}$$

Hence we find the linear subspaces corresponding to each kind of Chinese word images. The centroids of each word in PCA subspaces are given by

$$P_i = [e_{k,1}\ \ e_{k,2}\ \ \cdots e_{k,N_l}\ ]^T \tag{3}$$

$$C_{i,k} = \frac{1}{J}\sum_{j=1}^{J} P_i x_{i,j,k} \tag{4}$$

$C_{i,k}$ indicates the feature vector of the k-th word written by the i-th person.

In order to find a set of characteristic words of one writer that best identify his writing style, we calculate the average distance of feature vector $C_{i,k}$ of one word written by a person with that $C_{i,k'}$ of the same word written by the others.

$$d_{i,k} = \frac{1}{I}\sum_{k'=1}^{I}\left\| C_{i,k} - C_{i,k'} \right\| \tag{5}$$

For each person, we obtain k inter-class distances and the sorted sequence: $d_{i,1} \geq d_{i,2} \geq \cdots\cdots \geq d_{i,k}$. M coefficients are selected to represent the person's handwriting style. Each coefficient corresponds to one characteristic word of that person.

3.4 Fusion approach to writer identification using the characteristic words

Given an unknown handwriting text, a discrimination function is put forward below to classify the text to a certain writer.

$$i = \arg\min_{l} g_l(C'_{l,k}, C_{l,k}) \tag{6}$$

$g_l(C'_{l,k}, C_{l,k})$ is the distance between the unknown text and the text recorded in the database and is defined as follows:

$$g_l(C'_{l,k}, C_{l,k}) = \sum_{k=1}^{M}\omega_k \left\| C'l,k - C_{l,k} \right\|^2 \tag{7}$$

When making comparison, one set of characteristic words of the known text in the database was selected. We project the same set of words from the unknown text onto the corresponding subspaces that obtained in the training phase.The fusion method we bring forward can ensure the most efficiency for determining the unknown writer's identification. Because each characteristic word is endowed with a coefficient ω_k .

$$\omega_k = d_{l,k} \Big/ \sum_{k=1}^{M} d_{l,k} \tag{8}$$

The larger the coefficient ω_k is, the more distinctive writing style of this word can be best exhibited.

4. Experimental Results

Several experiments were carried out to test the validity of our algorithm. Each writer's handwriting texts are

divided into two parts and each part covers 5 texts, one for training, the other for testing.

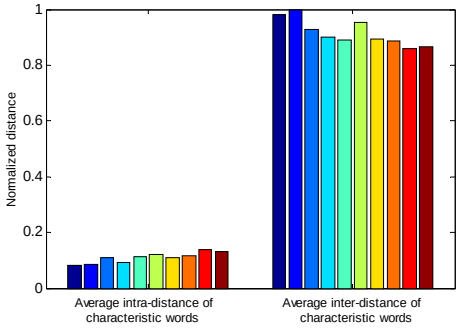

Figure 2. Normalized intra- and inter-class distances of 10 characteristic words

Figure 2 shows the normalized intra- and inter-class distance of 10 characteristic words. We can see from the figure that average intra-class distances of characteristic words are much smaller when compared with the average inter-class distances. So the characteristic words representation is a good way to show the difference of handwriting style. Such results prove the feasibility of the novel algorithm.

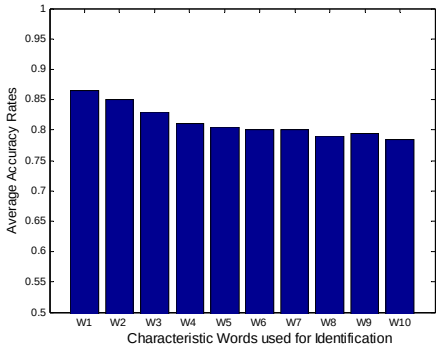

Figure 3. Result of the experiment using only one characteristic word for identification

Figure 3 gives the results of the experiment using only one characteristic word for identification. Each characteristic word here is shown by its corresponding coefficient. We can conclude from the results that when the more distinctive word is selected for identification, the better performance we can achieve. In the figure, W1 indicates the characteristic word with the largest coefficient; W2 indicates the second largest coefficient, and so on.

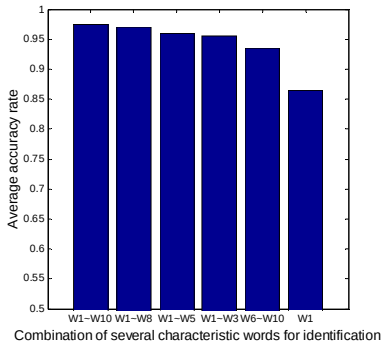

Figure 4. Result of the experiment using multi-characteristic words for identification

Usually one word cannot fully represent one person's writing style, so we combine a few more words to improve the accuracy of identification. Figure 4 shows the results of experiments with the combination of several characteristic words. Obviously using 10 characteristic words in the handwriting database can achieve the best result of 97.5% and

combining several words used for identification is much better than using only one word.

The above experiment results are all made under the gray-scale handwriting images. In contrast, we also give the experiment results under the binary images. Table 1 shows ratio of the normalized intra- and inter-class distances of characteristic words. From those, we can see that the ratios of intra- and inter-class distances of gray-scale images are smaller than those of binary images. So using gray-scale images can improve the performance of handwriting identification.

Tables 2 and 3 show the average accuracy rate of experiments using only one characteristic word contrast against that of using multi-characteristic words. Because the gray-scale images contain a significant amount of information reflecting the person's writing style conductive to the identification, so in our algorithm, we select gray-scale handwriting images. The results show its validity.

Table1: the ratio of intra- and inter-class distances

	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10
Gray-scale Images	0.083	0.084	0.118	0.102	0.127	0.127	0.129	0.131	0.158	0.159
Binary Images	0.085	0.092	0.117	0.128	0.131	0.139	0.141	0.152	0.155	0.163

Table2: the average accuracy rate of experiments using only one characteristic word.

	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10
Gray-scale Images	0.865	0.850	0.830	0.811	0.805	0.80	0.80	0.79	0.795	0.785
Binary Images	0.850	0.825	0.819	0.803	0.791	0.770	0.773	0.766	0.772	0.770

Table3: the average accuracy of experiments using multi-characteristic words.

	W1~W10	W1~W8	W1~W5	W1~W3	W5~W10
Gray-scale Images	0.975	0.970	0.960	0.955	0.935
Binary Images	0.962	0.960	0.955	0.951	0.938

5. CONCLUSIONS

In this paper, we apply PCA based statistical analysis to personal handwriting identification in gray-scale images and investigated the feasibility of this method based on the handwriting database containing 40 writers. Different from other methods, for one writer's handwriting, we only select a few characteristic words to indicate his writing style in the training phase. To judge an unknown handwriting text in the test phase, a fusion method was proposed to improve the performance of the recognition. Experimental results have shown the effectiveness of the new algorithm.

In practical applications, the proposed PCA method can improve the performance of writer identification by the following two aspects. First, the characteristic words of each writer can be obtained automatically and implicitly during the registration phase. Even the writer himself may not know which are his characteristic words. So it is very secure. Second, applying fusion of multiple characteristic words, which are content independent, to writer identification will improve the accuracy, comparing with content-fixed handwriting identification, such as signature verification. Third, gray-scale handwriting images containing pressure information are conducive to improvement of identification performance.

To sum up, all of the above demonstrate that combining characteristic words based on PCA is an efficient way for personal handwriting identification.

6. ACKNOWLEDGEMENT

The authors would like to thank Susan E. George for offering the dynamic handwriting data collection equipment.

7. REFERENCE

1. R. Plamond, G. Lorette, "Automatic Signature Verification and Writer Identification-the State of Art", *Pattern Recognition*, vol.22, no.2, pp.107-131, 1989.
2. H. E. S. Said, T. N. Tan and K. D. Baker, "Personal Identification Based on Handwriting", *Pattern Recognition*, vol.33, no.1, pp.149-160, 2000.
3. Rejean Plamondon and Sargur N. Srihari, "On-Line and Off-Line Handwriting Recognition: A Comprehensive

Survey”, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.22, no.1 January, 2000.

4. H. Bunke, Dr. U.-V. Marti, R. Messerli, “Writer Identification Using Text Line Based Features”, *Proceedings of 6th ICDAR '01*,
5. Yasushi Yamazaki and Naohisa Komatsu, “A proposal for Text-Indicated Writer Verification Method”. *Proceedings of ICDAR’97*
6. E.N. Zois, and V. Anastassopouls, “Fusion of Correlated Decisions for Writer Verification, Pattern Recognition”, *vol. 32, NO. 10, pp. 1821-1823, 1999*
7. S. Cha and S. N. Srihari, “Multiple Feature Integration for Writer Verification”, *the Proceedings of 7th IWFHR2000, Amsterdam, Netherlands, September 2000, p 333-342*
8. Mami KIKUCHI and Norio AKAMATSU, “Development of Speedy and High Sensitive Pen System for Writing Pressure and Writer Identification”. *Proceedings of 6th ICDAR '01*,