# Online Text-Independent Writer Identification Based on Stroke's Probability Distribution Function

Bangyu Li, Zhenan Sun, and Tieniu Tan

Center for Biometrics and Security Research, National Lab of Pattern Recognition,
Institute of Automation, Chinese Academy of Science, Beijing, P.R. China
{byli,znsun,tnt}@nlpr.ia.ac.cn

**Abstract.** This paper introduces a novel method for online writer identification. Traditional methods make use of the distribution of directions in handwritten traces. The novelty of this paper comes from 1)We propose a text-independent writer identification that uses handwriting stroke's probability distribution function (SPDF) as writer features; 2)We extract four dynamic features to characterize writer individuality; 3)We develop new distance measurement and combine dynamic features in reducing the number of characters required for online text-independent writer identification. In particular, we performed comparative studies of different similarity measures in our experiments. Experiments were conducted on the NLPR handwriting database involving 55 persons. The results show that the new method can improve the identification accuracy and reduce the number of characters required.

**Keywords:** text-independent, writer identification, stroke's probability distribution function, dynamic features.

## 1 Introduction

Writer identification is the task of determining the writer of a sample handwriting [1]. Surveys covering work in automatic writer identification and signature verification until 1993 are given in [2][1]. Traditionally, research into writer identification has been focused on two streams: offline and online writer identification [3]. Scarce research results can be found in the online direction, in particular, personal identification using online Chinese handwriting. So far, two different kinds of methods find their ways in handwriting authentication: text-dependent way and text-independent way. Text-independent methods have prominent advantages over text-dependent cases in specific conditions. Text with the same content is not required for the training samples and the testing ones, and the natural handwriting in a wide range can be dealt with [4]. It is difficult for the imposter to imitate the handwriting of others. Therefore, text-independent methods have gained more and more attention in recent years.

Recently, a number of new approaches to writer identification have been proposed, such as those using connected-component contours and edge-based features [5], using dynamic features on strokes [6][7][8] or on hinge angles [9][10],

S.-W. Lee and S.Z. Li (Eds.): ICB 2007, LNCS 4642, pp. 201–210, 2007.

using texture feature[3] [11][12], using innovative binarised features [13], using moment feature [14], etc. Furthermore, some classical methods for character recognition are helpful, such as statistical character structure modeling [10], dynamic time warping (DTW) or cluster generative statistical dynamic time warping (CSDTW) [4] and so on. In these methods, dynamic or static features are widely adopted to discriminate the handwriting. They have been proved to be feasible, but some disadvantages still exist.

The proposed algorithm is implemented based on the strokes of the characters, for the Chinese characters are composed of multiple interlaced strokes in most cases. In this paper, with the information extracted from the strokes, a comprehensive analysis of the writing style is obtained. We propose a novel approach for online writer identification based on the stroke's probability distribution function.

The paper is structured as follows. The next section presents previous works and motivation. The dynamic features of strokes are extracted from handwriting in Section 3. In Section 4 we describe the writer identification algorithm with distance between histograms. The results of our experiments and discussion are presented in Section 5. Finally, in Section 6 we conclude the paper.

## 2 Previous Work and Motivation

### 2.1 Previous Work

We adopted the traditional framework of pattern recognition to our handwriting identification problem, as Fig. 1 shows [6].

On-line handwriting samples are obtained as a sequence of parameterized dots, inside which involves coordinates, relative time, pressure and pen inclination information. The raw data are tailored to fit in with the basic requirement

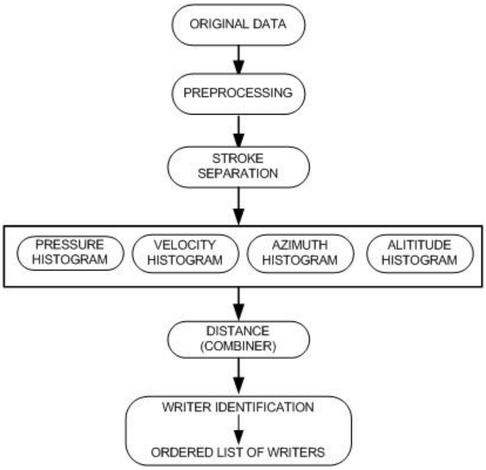


**Fig. 1.** Schematic diagram of the proposed algorithm

of subsequent process. Here only the fake writing movements are eliminated. Afterwards, the separation of strokes is implemented in three steps, namely 1) line separation, 2) connection separation and 3) stroke separation. The last step of preprocess is stroke type evaluation, where strokes are allocated to predefined stroke types. According to the structural constituents of the Chinese characters, 12 primary stroke types are selected, as Fig. 2 shows, and the arrow below each stroke indicates the writing direction of the stroke. Details may be found in [6].

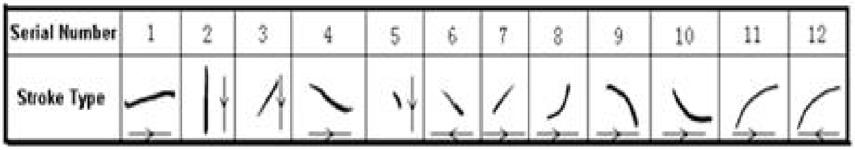


**Fig. 2.** The 12 Primary Stroke Types

### 2.2 Motivation

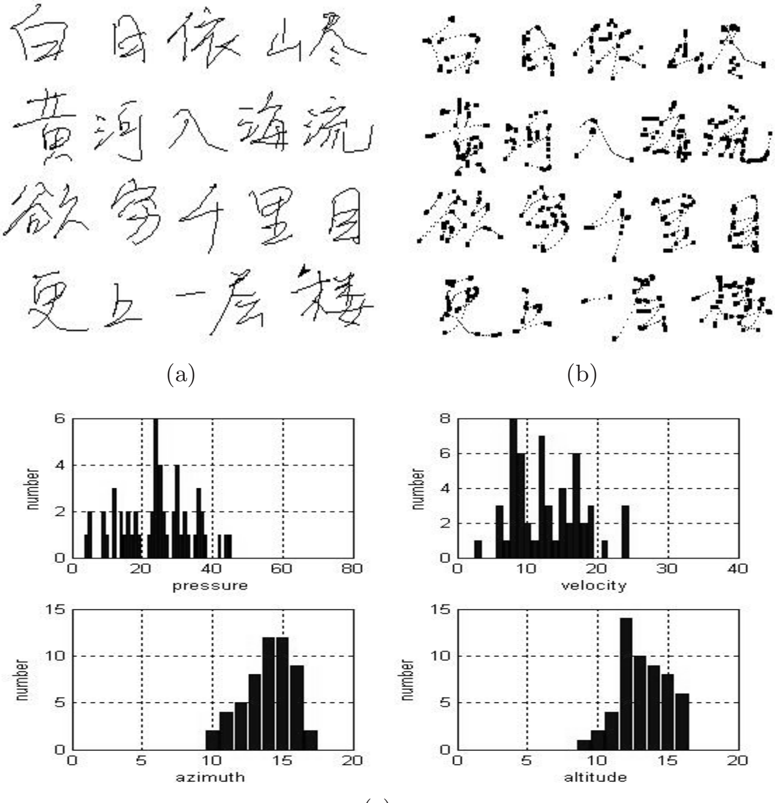
Online text independent writer identification usually requires the use of statistical features computed from a large quantity of data to avoid anomalies due to specific text. Our previous method builds Gaussian models that each model could be represented by its mean value and variance[6], which is lack of sufficient information to characterize dynamic attribute of strokes in handwriting.

We use the stroke's probability distribution function (SPDF) of four dynamic features to characterize writer individuality. The smallest units forming a Chinese character are strokes, and this method aims to capture more dynamic aspects of the writing behavior of an individual. Writing different Chinese characters will lead to the difference of strokes' statistical information and the emotional and physical state of writers will have influence on its dynamic features. However, this method uses statistical features of primary stroke types and four dynamic features are divided into small units that characterize more detailed writer individuality. The stroke's occurrence of PDF is a discriminatory feature among different writers. Moreover, the dynamic features can be computed very fast using histogram representation. So we decide to choose SPDF to model dynamic features.

## 3 Feature Extraction

The feature extraction procedure is described in this section. It has long been known from online handwriting research that the distribution of directions in handwritten traces, as a polar plot, yields useful information for writer identification [15][16][17].

We use probability distribution functions extracted from strokes in handwriting samples to characterize writer individuality in an online text-independent manner. The raw data collected from the tablet includes many fake strokes and noises, as Fig. 3(a) shows. After the preprocessing of wiping off fake strokes, the separation of strokes is implemented, as Fig. 3(b) shows. In our analysis, we will consider a number of features that we have designed (f1, f2, f3, f4) for writer identification. Strokes well allocated help to simplify further analysis; features are extracted from the main 12 types of strokes in our method. We select distribution of the pressure, the velocity, the altitude and azimuth of $i = 12$ primary stroke types for the representations of the writing gestures and movements. We define four features (f1, f2, f3, f4) of stroke's probability distribution function. For pressure feature f1, the number of histogram bins spanning the interval 0-1024 was set $N = 64$ to through experimentation: 16/bin gives a sufficiently detailed and sufficiently robust description of handwriting to be used in writer identification. For velocity feature f2, by experiment, the number of histogram bins spanning the interval 0-300 was set to $N = 30$, so each bin has 10. The



(a) (b)

(c)

**Fig. 3.** (a)Original online text. (b)Stroke separation, where the black dots represent the starting and end points of a segment. (c)Pressure,velocity, azimuth and altitude histograms of Stroke type 1.

azimuth denotes the angle between the projection of the pen on the tablet and the vertical axis of the digital tablet, the altitude denotes the angle between the pen and its projection on the tablet, and synthesis of the two angles fixed the pen-grasping gesture of the writer. Because altitude ranges from $0^0$ to $90^0$ and azimuth ranges from $0^0$ to $180^0$, we choose $5^0$ and $10^0$ as a bin, respectively, as Figure 3(c) shows.

An overview of all the features used in our study is given in Table 1. In our analysis, we will consider four features based on strokes that we have designed and used for writer identification.

**Table 1.** Overview of the considered features

| | Feature | Explanation | N dimensions |
|---|---|---|---|
| f1 | $h_{1i}(A)$ | Pressure-histogram | 12*64 |
| f1 | $h_{2i}(A)$ | Velocity-histogram | 12*30 |
| f3 | $h_{3i}(A)$ | Azimuth-histogram | 12*18 |
| f4 | $h_{4i}(A)$ | Altitude-histogram | 12*18 |

## 4 Identification

Feature vectors of the testing data are extracted as well as the training data. The identification of writers based on given feature vectors is a typical recognition problem.

Writer identification is performed using nearest-neighbor classification in a "leave-one-out" strategy. For a query sample q, the distances to all the other samples $i \neq q$ are computed using a selected feature, and then all the samples are ordered in a sorted hit list with increasing distance to the query q. Ideally the first ranked sample should be the pair sample produced by the same writer. In our feature combination scheme, the five distances between any two handwritten samples is computed by weighted Sum-Rule of the dynamic features.

Most of the distance measures presented in the literature consider the overlap or intersection between two histograms as function of the distance value, but they do not take into account the similarity between the non-overlapping parts of the two histograms.

We choose new distances between histograms defined on a structure called signature, which is a lossless representation of histograms, the main advantage is that there is an important time-complexity reduction. Moreover, the type of the elements of the sets that the histograms represent are ordinal, nominal and modulo [18].

## 5 Experimental Results and Discussion

Our experiments are based on the NLPR handwriting database. In this database, samples are collected using a Wacom Intuos2 tablet. The tablet has 1024 levels of pressure sensitivity, and its maximum data rate is 200pps. 55 persons were required to write three different pages including approximately 600 Chinese characters altogether. The content of the handwriting is chosen by writers freely, so the database collected by this manner is text-independent handwriting database.

When we carry out the experiments with the method using dynamic features, two pages samples of each person are taken out for training and the other one (approximately 160 characters) is used for testing. We use $x^2$ distance, Euclidean distance and three distances [18] as a distance measure for SPDF features.

We are interested in a comparative performance analysis of the different features and in the improvements in performance obtained by combining multiple features. First we shall consider the individual features and then their combinations.

### 5.1 Performances of Individual Features

Fig. 4 gives the writer identification performance of the individual features considered in the present study. While there are important different in performance among the different features, the best performer is the velocity feature followed by the pressure feature. We select samples with 2 lines, 3 lines, 4 lines, 5 lines, 6 lines and 7 lines of characters for each time. The results are shown in Fig. 4.

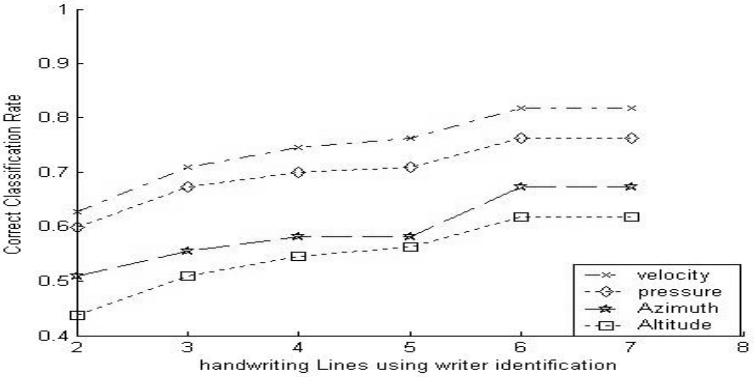


**Fig. 4.** The identification results for using individual feature

### 5.2 Performances of Feature Combinations

The features considered in previous experiments capture different aspects of handwriting individuality, while our features are not completely orthogonal, combining multiple feature proves to be beneficial. In this paper, we use min-max rule for score normalization [19]. Given a set of matching scores $\{S_k\}$, $k = 1, 2, \cdots, n$, the normalized scores are given by

$$S_k^n = \frac{S_k - min}{max - min} \tag{1}$$

min and max are the minimum and maximum values estimated from $\{S_k\}$. The weighted sum rule is used to combine scores from each dynamic feature for text-independent writer identification. Let $\{S_1^j, S_2^j, \cdots, S_n^j\}$ be the matching scores for each dynamic feature, $j = 1, 2, \cdots, c$. Here, $n$ is the number of samples, and $c$ is the number of dynamic features. For the $i$th sample, the combined score can be computed as

$$S_i = \sum_{j=1}^{c} w^j * S_i^j \tag{2}$$

$w^j$ is an indicator contribution coefficient defined as the confidence of each dynamic features. The confidence of each classifier is computed by the training result(the first experiment) as

$$w^j = \frac{FIR_j{}^{-1}}{\sum_{i=1}^{c} FIR_i{}^{-1}} \tag{3}$$

FIR is the False Identification Rate. Combination of dynamic features by five distances is performance in our method. We compare five distance in fusing features. The results are shown in Fig. 5.

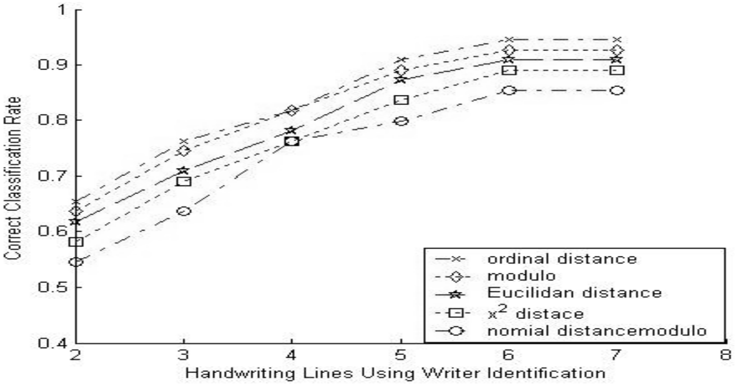


**Fig. 5.** The identification results for using feature combinations

For testing effectiveness of our method, we choose SVC 2004 database [20] and each signature as a stroke. The performance of feature combinations for writer identification on SVC 2004 database. We choose 40 writers and each writer has 20 genuine signatures for our experiments, we apply feature combinations for writer identification, we can get results shown in Table 2.

**Table 2.** EER performance in our experiments

| EER | Euclideanm dist | $x^2$dist | Nominal dist | Ordinal dist | Modulo dist |
|---|---|---|---|---|---|
| SVC database | 11.4% | 11.8% | 15% | 9.7% | 10.4% |

The fourth set of experiments were performed to determine the effectiveness of feature extraction to classify handwritten data from strokes. Each stroke were extracted from one page (approx.8 lines) and the four dynamic features of each stroke was combined by the weighted sum rule. Table 3 shows the accuracy of each type of strokes. For other type of strokes, only 40% accuracy could be achieved, as the number of strokes extracted from handwriting is not so little that do contain much individuality information.

**Table 3.** Strokes Vs Accuracy

| Type of Stroke | top-1 Accuracy | top-5 Accuracy |
|---|---|---|
| The horizontal(type 1) | 70 | 92 |
| The Verticcal(type 2) | 68 | 90 |
| The Left-falling Stroke(type 3) | 60 | 73 |
| The Right-falling Stroke(type 4) | 63 | 74 |

### 5.3 Discussion

From experimental results we can find that: First, as Fig. 4 shows, it is important to observe that the feature f1 and f2 perform much better than the others, and the performance of the proposed method is also tested with increasing characters, where the best result is gained when the number of writing lines is beyond 6, including approximate 100 characters. The main reason for this result is that the range of azimuth and altitude are changing little in raw data, it is not sufficiently robust description of handwriting to be used in writer identification.

Second, as Fig. 5 shows, we can see that the results of combination are obviously better than individual feature. It proved our idea that combining the dynamic features could improve the result of identification. Further more, from the results, we can see that when we use samples with 6 lines (100 characters on average) of characters, it can achieve the accuracy of using samples with 7 lines (more than 120 characters on average) in experiments adopting individual dynamic features. The performance of ordinal distance is especially satisfactory. It shows that the combination method not only can improve the identification accuracy, but also reduces the amount of characters (lines) required in handwriting.

At last, as Table 2 shows, the ordinal distance performs better than the others. In SVC 2004 database, we find that distribution of pressure and velocity features can effectively describe the characteristics of the writer handwriting. The writer identification result is about 10% in terms of EER. The notable point is that the based on SPDF's feature can effectively describe the handwriting characteristics. and Table 3 shows, most of the shape primitive of Chinese character were straight lines. We can improve our accuracy as number of strokes increases, in future we can use straight lines to improve on the accuracy.

## 6 Conclusion

In this paper, we have proposed a novel method for online text-independent writer identification. In our method we not only adopt individual features based on SPDF for writer identification, but also choose feature combinations in some distance measures to improve identification accuracy. Experimental results show the dynamic features of strokes is a very stable personal characteristic and the most prominent attribute of online handwriting that reveals individual writing style and the dynamic features can be computed very fast using histogram representation with the additional advantage, the new method can improve the identification accuracy and reduce the number of characters required.

**Acknowledgments.** This work is funded by research grants from the National Basic Research Program (Grant No. 2004CB318110), the Natural Science Foundation of China (Grant No. 60335010, 60121302, 60275003, 60332010, 6982510560605008) and the Chinese Academy of Sciences.

## References

1. Plamondon, R., Lorette, G.: Automatic signature verification and writer identification -the state of the art. Pattern Recognition, 107–131 (1989)
2. leclerc, F., Plamondon, R.: Automatic signature verification:the state of the art 1989-1993. International Journal of Pattern Recognition and Artificial Intelligence, 643–660 (1994)
3. Tan, T.N., Said, H., Baker, K.D.: Personal identification based on handwriting. Pattern Recognition, 149–160 (2000)
4. Bahlmann, C., Burkhardt, H.: The writer independent online handwriting recognition system frog on hand and cluster generative statistical dynamic time warping. IEEE Transactions on Pattern Analysis and Machine Intelligenc 1, 299–310 (2004)
5. Schomaker, L., Bulacu, M.: Automatic writer identification using connected-component contours and edge-based features of uppercase western script. IEEE Transactions on Pattern Analysis and Machine Intelligence 6, 787–798 (2004)
6. Wang, Y., Yu, K., Tan, T.: Writer identification using dynamic features. International Conference on Biometrics, 512–518 (July 2004)
7. Messerli, R., Marti, U.V., Bunke, H.: Writer identification using text line based features. In: Internal Conference on Document Anaysis and Recognition, pp. 101–105 (2001)
8. Schomaker, L.R.B., Plamondon, R.: The relation between pen force and pen-point kinematics in handwriting. Biological Cybernetics, 277–289 (1990)
9. Methasate, I., Sae-Tang, S.: On-line thai handwriting character recognition using stroke segmentation with hmm. In: International conference on Applied informatis-Artificial Intelligence and Applications, pp. 59–62 (2002)
10. Kim, I.-J., Kim, J.-H.: Statistical character structure modeling and its application to handwritten chinese character recognition. Biological Cybernetics, 1422–1436 (2003)
11. Wang, Y., Zhu, Y., Tan, T.: Biometric personal identification based on handwriting. International Conference on Pattern Recognition, 801–804 (2001)

12. Tan, T.N.: Texture feature extraction via cortical channel modeling. In: Proc.11th IAPR Inter. Conf. Pattern Recognition, pp. 607–610 (1992)
13. Leedham, G., Chachra, S.: Writer identification using innovative binarised features of handwritten numerals. In: Internal Conference on Document Anaysis and Recognition, pp. 413–417 (2003)
14. Mertzios, B.G., Tsirikolias, K.: Statistical pattern recognition using efficient two-dimensional moments with applications to character recognition. pattern recognition, 877–882 (1993)
15. maarse, F., Thomassen, A.: Produced and perceived writing slant:differences between up and down strokes. Acta psychologica 3, 131–147 (1983)
16. Schomaker, L., Maarse, F., Teulings, H.-L.: Automatic identification of writers. In: Human-computer interaction: Psychonomic aspects, pp. 353–360. Springer, Heidelberg (1988)
17. Crettez, J.-P.: A set of handwriting families:style recognition. In: Internal Conference on Document Anaysis and Recognition, pp. 489–494 (1995)
18. Serratosa, F., Sanfeliu, A.: Signatures versus histograms: Defintions, distances and algorithms. Pattern Recgnition, 921–934 (2006)
19. Jain, A.K., Nandakumar, K., Ross, A.: Score normalization in multimodal biometric systems. Pattern Recognition 12, 2270–2285 (2005)
20. Yeung, D.Y, Chang, H., Xiong, Y., George, S., Kashi, R., Matsumoto, T., Rigoll, G.: Svc 2004: First international signature verification competition. In: Zhang, D., Jain, A.K. (eds.) ICBA 2004. LNCS, vol. 3072, pp. 16–22. Springer, Heidelberg (2004)