# Writer Identification Method using Inter and Intra-Stroke Information

Jungpil Shin, Zhaofeng Liu
School of Computer Science and Engineering,
The University of Aizu, Fukushima, Japan
E-mail: jpshin@u-aizu.ac.jp

***Abstract*— Most research in the field of writer identification currently uses positional data to identify the writer. But there is no writer identification method that is suitable for practical use so far. So, we newly devise two algorithms to raise accuracy rate and find some parameters that have beneficial effects on writer identification. One algorithm is "Block Type Model" which is a method to analyze positional relation and shapes of Chinese characters. The other algorithm is "Hidden feature analysis" which is an algorithm that identifies the writer using multi-parameters with all the strokes. As a result, any eight Chinese characters were enough to achieve over 99.9% accuracy rate. Additionally, we discovered that there are personal attributes in the speed of each stroke. With the new algorithm, we raised accuracy rate and found out that parameters have beneficial effects on writer identification.**

***Keywords- Writer Identification; Stroke Information***

## I. INTRODUCTION

Hand-written characters contain unique characteristics of people. Recently, what is most desired in the field of handwriting analysis is a proposal of a method or algorithm with an accuracy which is higher than or equivalent to the fingerprint authentication system or the vein authentication system.

To identify the characteristics of the writer, many algorithms have been proposed, but their identification success rate is not at the level of practical use. The biggest reason is that the amount of information obtained from the position of dots and lines is not sufficient to identify the writer.

The methods to extract characteristics of the writer from the limited information can be divided into two main categories. One is known and described as an "Intra" type. This type verifies the shape of a line of characters or the position among radicals. The other is called as an "Inter" type, which verifies the shape of every part of characters (e.g., distance between each stroke). Using the two methods is expected to bring higher accuracy.

Edge-Based Directional Features, connected-component contours and other methods have been invented in previous works. Off-line means the samples were collected from papers or by some analog method. On-line means the samples were collected from tablets or by some digital method. Number of characters means how many characters were necessary for matching.

Many algorithms analyze the position information of lines or points, and extract individual features. However, there are few methods that analyze the pen pressure, stroke speed and an inclination angle of a pen. Moreover, there are few databases which contain huge amount of location data. Therefore, we newly devised two algorithms. One is an "Inter" type algorithm which uses the relationship among block such as radicals. The other is an "Intra" type algorithm with position data, pen pressure, speed data, and an inclination angle of a pen, an angular difference between paper and pen. Then, we combine these two methods and do an experiment using the database that we have been creating for over five years. The database currently contains 110 kinds of Chinese characters, written by 48 writers of different nationalities, 10 times each, adding up to 52,800 characters in total. In this research, half of the characters were used to extract the characteristics of the writers, and the other half was used for an identification experiment. The results indicate that features extracted from eight Chinese characters were enough to identify the writer. In Section 2 we show the approaches for writer identification. In Section 3 we show the experiment and results. In Section 4 we show the discussions. In Section 5 we conclude the paper.

## II. WRITER IDENTIFICATION APPROACHES

In this section, we present an algorithm for writer identification using the Block Type Model, Hidden Feature Analysis.

### *A. Initialization*

To identify the writer who has written Chinese characters, we used a database that contains positional data, pen pressure, inclination angle of the pen, angular difference between paper and pen, and pen speed. When we collected character data, we had no idea how to standardize the size of characters. In this research, we have to standardize the size of characters for matching. First we computed the average size of all the characters in the database for Dynamic Time Warping (DTW) matching. Then, we used it as the standard size of characters. However, since there are vertically long and horizontally long types in Chinese characters, expansion and reduction cannot be simply performed in every direction by the same ratio. Therefore, we distinguished the characters of a vertically long or horizontally long type. We divided the Chinese characters into two or more blocks: a radical block and another block. In the case of a Chinese character with two or more radicals, we added the number of blocks as radical block.

 ISBN: 978-1-61208-416-9 

TABLE I. RESULTS OF BLOCK TYPE MODEL (%)

| No. of Char. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Block | 62.94 | 80.02 | 88.52 | 91.95 | 94.83 | 96.41 | 97.48 | 98.15 | 98.62 | 98.96 |

We further classified them according to the direction of vectors between the two most distant blocks. If the vector pointed sideways, we posited it as a horizontally long type character. If the vector was vertical, we posited it as a vertically long type. In particular, we used the expressions below for making judgements.

TABLE II. HIDDEN FEATURE ANALYSIS WITHOUT WEIGHT

| No. of char | Pressure | Theta | Phi |
|---|---|---|---|
| 1 | 55.45 | 33.18 | 33.60 |
| 2 | 73.54 | 42.02 | 45.40 |
| 3 | 81.98 | 47.89 | 52.56 |
| 4 | 86.93 | 51.49 | 57.35 |
| 5 | 89.95 | 54.05 | 61.45 |
| 6 | 91.90 | 56.32 | 64.66 |
| 7 | 93.41 | 58.29 | 67.18 |
| 8 | 94.46 | 59.44 | 69.51 |
| 9 | 95.40 | 60.56 | 71.60 |
| 10 | 96.02 | 61.63 | 73.24 |

The relation between the lengthwise direction of the interior of characters can be an important factor for a vertically long type of character, and the relation between the crosswise direction of the interior of characters can be an important factor for a horizontally long type of character. So, we resized characters according to the type of character (horizontally long or vertically long). If the character was vertically long, we resized it with the ratio of the lengthwise direction of the standard. If the character was horizontally long, we resized it with the ratio of the crosswise direction of the standard; we did not change the vertical ratio of vertically long type characters or the horizontal ratio of horizontally long type characters in Figure 1.

### B. Block Type Model

As previously noted, our research attempts to combine an inter-type algorithm, the Block Type Model, with an intra-type algorithm, Hidden Feature Analysis.

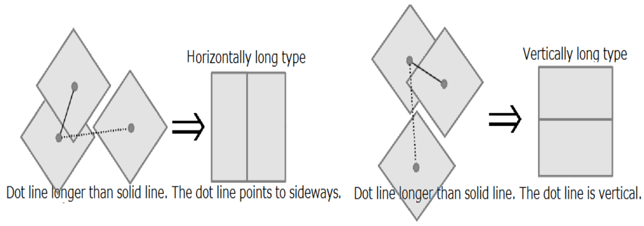


Figure 1. Graphic explanation about character types

1. A character is divided into some blocks.
2. The points that contain the maximum X-coordinate, the minimum X-coordinate, the maximum Y-coordinate, and the minimum Y-coordinate of each block are recorded.
3. The average of points of the same kind of character written by the same writers is computed. Then it is defined as a class of writer (e.g., Writer A's character B class)
4. The points of characters and points of classes are compared according to the Euclidean distance and weight (it should be defined as an approximate value). Then the class of the best approximation of the character is identified.
5. The class with the lowest approximate value is determined as the writer's class.

Weights are decided with the rule stated below.

Weight = Average of variance of each point / Variance of each point.

The approximate value is defined by the equation below number of characters.

$$A=E*W$$

where, A, E, and W is approximate value, Euclidean distance between each pair of points, and weight of each point, respectively.

In other words, higher variance means a lower approximate value. To put it another way, lower variance means a higher approximate value. A graphic explanation is given in Figure 2. In Figure 2, points A, B, C, D, and so on mean the points that have the maximum X-coordinate and Y-coordinate and minimum X-coordinate and Y-coordinate.

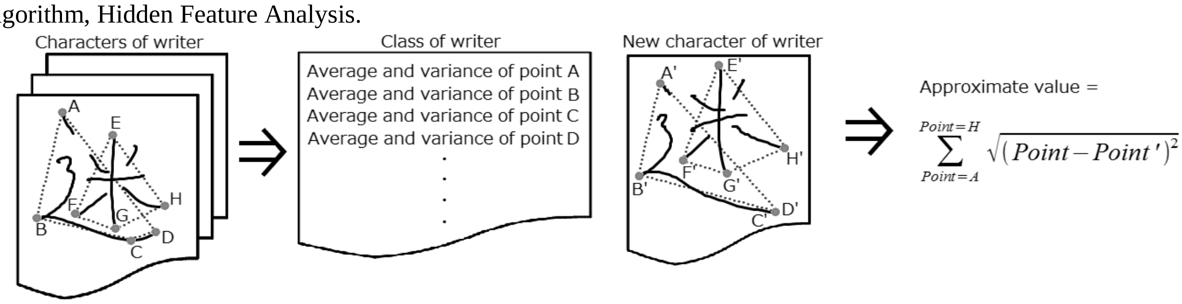


Figure 2. Graphic explanation about class

TABLE III. WEIGHTS OF HIDDEN FEATURE ANALYSIS.

| | Average of accuracy of each parameter (%) | Weight |
|---|---|---|
| Pressure | 85.91 | 1.18 |
| Theta | 52.49 | 0.72 |
| Phi | 59.67 | 0.82 |
| Speed | 93.68 | 1.28 |
| Average | 72.94 | |

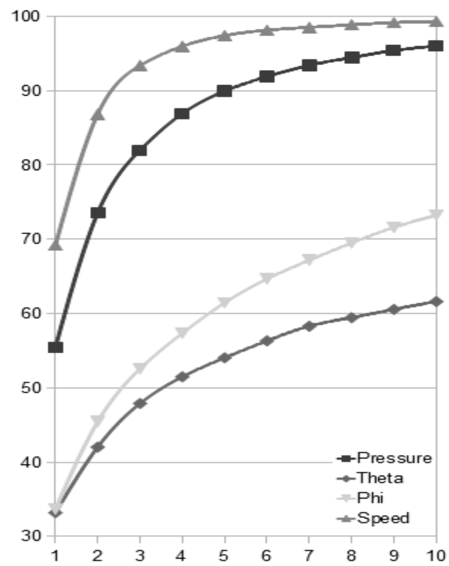


Figure 3. Results of Hidden feature analysis without weights (%).

We only show the results of the Block Type Model (Table 1), "No. of char." means the number of Chinese characters used for matching.

This experiment was conducted using 110 kinds of Chinese characters written by 48 writers. One writer wrote 10 times each character. Namely, we collected 52800 characters in total. This experiment was conducted with 1 to 10 characters chosen randomly from the database, calculating the distance between the class and the character for each character. We used the sum total of the approximate values to identify the writer. Then, we repeated this 1000 times. We obtain the results of the average.

### C. Hidden feature Analysis

This is an algorithm that identifies the writer using the average of pen pressure, the inclination angle of the pen, the angular difference between paper and pen, and the speed and range of these parameters in each stroke. The Hidden Feature Analysis was carried out as below.

1. Characters were divided into strokes.
2. The average of pen pressure, inclination angle of the pen, angular difference between paper and pen, and speed and range of these parameters in each stroke were computed. Then, we defined the character as a class of writer (e.g., Writer A's character B class)
3. We compare the written characters with class. (It was defined as an approximate value.)

Then, we identified the class of the best approximation to the character. (The class with the lowest approximate value was determined to be the writer's class.)

Graphic explanation is given in Figure 1. The results of experiment with an individual parameter without its weight are found in Table 2.

In Figure 3, "pressure" and "speed" mean the pen pressure and speed of each stroke respectively. Theta is the inclination angle of the pen. "Phi" is the angular difference between paper and pen. The horizontal axis corresponds to the number of characters used for the analysis and the vertical axis corresponds to the correct rate.

The result shows that there is a difference in matching performance among the parameters. Then the average accuracy of each parameter is computed (Table 3).

We defined the average of the overall parameters as the standard and used the rule below to decide the weights.

Average of accuracy of each parameter /
Average of accuracy of all parameter

The weight of each parameter is given in Table 3. In other words, we used the weight to improve the influence of the parameters that were performing well and also to lessen the influence of the parameters that were not performing so well.

The results of the combination of the Block Type Model and Hidden Feature Analysis are given later. Here we only show the results of the Hidden Feature Analysis (Table 5). "Hidden feature" means the accuracy rate without the weight. "Hidden feature with weight" means the accuracy with the weight.

This experiment was conducted with 1 to 10 characters which were chosen randomly from the database, calculating the distance between the class and the character for each character. We used the sum total of the approximate values to identify the writer. Then, we repeated this 1000 times. We obtained the results of the average.

TABLE IV. RESULTS OF HIDDEN FEATURE ANALYSIS (%).

| No. of char | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 4-Axis | 87.79 | 95.1 | 97.19 | 98.25 | 98.85 | 99.24 | 99.45 | 99.63 | 99.73 | 99.82 |
| 4-Axis with weight | 88.80 | 95.68 | 97.61 | 98.56 | 99.12 | 99.48 | 99.64 | 99.77 | 99.83 | 99.89 |

## III. MAIN EXPERIMENT AND RESULTS

We describe the results of two experiments: (1) an inter-type Block Type Model and (2) an intra-type Hidden Feature Analysis. Now, we combine these two methods and carry out the experiment. In particular, we summed up the two methods' approximate values and appointed the class that had the lowest approximate value as the writer's class.

This experiment was conducted using 110 kinds of Chinese characters written by 48 writers. Each writer wrote each character 10 times. Namely, we collected 52800 characters in total.

Comprehensive results are given in Table 5. This experiment was conducted with 1 to 10 characters chosen randomly from the database, and the distance between the class and the character was calculated for each character. We used the sum total of the approximate values to identify the writer. Then, we repeated this 1000 times. We obtained the results of the average. Additionally, the ratio of approximate values of the two algorithms is set to 1:1.

There are some researches similar to our research [4] and [5]. Our database is smaller than that used in [5], but we can achieve a higher accuracy rate with a smaller number of characters. We can achieve 90% accuracy with one character, while [5] needed 50 words. Moreover, we can achieve a higher accuracy rate than [4] in the same situation. Their accuracy was 97.5%. But our maximum accuracy rate is 99.9%.

TABLE V. RESULTS OF COMBINATION OF BLOCK TYPE MODEL AND HIDDEN FEATURE ANALYSIS (%).

| No. of char | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Result of combination of Rhombus type model and 4-Axis analysis | 90.53 | 96.71 | 98.52 | 99.24 | 99.58 | 99.81 | 99.86 | 99.91 | 99.94 | 99.98 |

## IV. CONCLUSION

In this paper, we presented novel two algorithms to improve the accuracy rate and find some parameters that have beneficial effects on writer identification. The results indicate that four Chinese characters were enough to achieve an accuracy rate of 99% and eight Chinese characters were enough to achieve an accuracy rate of over 99.9%. Table 5 indicates that the accuracy rate of the two methods combined is higher than the results of the two individual approaches. This means that this method is quite useful for identifying the writer.

Our algorithm does not support one-to-many matching. This means that when we identify the writer, the written Chinese character must be the same kind as was written at the time of class generation. In the future, we hope to be able to generate a class that can support more Chinese characters with a small number of characters..

## REFERENCES


[1] S. N. Srihari, S.-H. Cha, H. Arora, and S. Lee, “IIndividuality of Handwriting,”, Journal of Forensic Sciences, 47(4), pp. 1-17, July 2002.

[2] E. Zois, “Morphological Waveform Coding for Writer Identification,” Electronics Laboratory, Physics Department, University of Patras, Greece, March. No.33, pp.385-398, 2000.

[3] C. Tomai, B, Zhang, and S. Srihari, “Discriminatory Power of Handwritten Words for Writer Recognition,” Intl. Conf. of Pattern Recognition (ICPR2004), Vol. 2, pp. 638-641, NY, 2004.

[4] Z. Long, W. Yunhong, and T. Tieniu, “Personal Handwriting Identification Based on PCA,” Proceedings of SPIE Second International Conference on Image and Graphics, pp 766-771, 2002.

[5] B. Li, Z. Sun, and T. Tan, “Hierarchical Shape Primitive Features for Online Text-independent Writer Identification,” 10th International Conference on Document Analysis and Recognition, ICDAR 2009, pp. 986-990, Barcelona, Spain, 26-29 July 2009.

[6] Y. Yan, C. Chen, W Deng and F. Yuan, “Chinese Handwriting Identification Based on Stable Spectral Feature of Texture Images,” International Journal of Intelligent Engineering and Systems, Vol. 2, No. 1, pp.17-22, 2009.

[7] B. Zhang, S. Srihari, and S. Lee, “Individuality of Handwritten Characters,” Seventh Int‘l Conf. Document Analysis and Recognition (ICDAR), pp. 1086 , Aug. 2003.

[8] N. Mohammed, E. AlKassis, and D. Al-Muslih, “Off-Line Text-Independent Arabic Writer Identification using Contour-Based Features,” Mir@cl Lab, FSEGS, University of Sfax, 2010.

[9] H. Said, T. Tan, and K Baker, “Personal identification based on handwriting,” Pattern Recognition, Vol 33, No 1, pp. 149–160 January 2000.

[10] L. Schomaker and M. Bulacu, “Automatic Writer Identification Using Connected-Component Contours and Edge-Based Features of Uppercase Western Script,” IEEE, Transaction on pattern analysis and machine intelligence, Vol. 26, No. 6 June. 2004.

[11] A. Bensefia, T. Paquet, and L. Heutte. , “A writer identification and verification system,” Pattern Recognition Letters, Vol 26, No. 13, pp. 2080–2092, October 2005.

[12] C. Siewkeng, T. Yonghaur, and V. Christian, “Online Text Independent Writer Identification Using Character Prototypes Distribution,” 6th International Conference on Information, Communications & Signal Processing, pp.1-5, 10-13, Dec. 2007, Utar, Malaysia.

[13] T. Matsuura, P.Thumwarin, and N. Kato, "On-live writer identification based on handwriting velocity and curvature of script," 23rd International Conf., on Image and Vision Computing. pp. 192-197, 26-28 Nov. 2008, New Zealand,.

[14] A. Schlapbach, M. Liwicki, and H. Bunke, "A Writer Identification System for On-line Whiteboard Data," Pattern Recognition, Vol 41, No 7, pp. 2381–2397, July 2008.