An Empirical Study of Text Factors and Their Effects on Chinese Writer Identification

Yu-Jie Xiong[1,2(✉)], Yue Lu[2], and Yan-Chun Cao[3]

[1] School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China
xiong@sues.edu.cn

[2] Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University, Shanghai 200241, China

[3] School of Public Administration, Faculty of Economics and Management, East China Normal University, Shanghai 200062, China

Abstract. In this paper, we analyze the relationship between the performance of the text-independent feature and text factors of the handwriting on Chinese writer identification. Text factors contain two types of information: the number of characters in both query and reference and the number of the same characters in both query and reference. We conclude that the performance increases when the query and reference contain more characters, and the minimum number of needed characters is 50. The number of the same characters in both query and reference has little influence on the identification when the number of characters is more than 50. The conclusions are verified by repeated writer identification tests with different amount of characters on the handwriting document pages.

Keywords: Chinese writer identification · Empirical study · Text factors · Text-independent

1 Introduction

Biometrics are becoming a key aspect of information security [1]. Conventional authentication techniques use a special product (i.e., key, ID card, etc.) or a unique information (i.e., password, etc.) to perform personal recognition. However, biometrics refers to individual recognition based on a person's physical or behavioral traits. Physiological and behavioral biometrics are two main branches of biometrics, and writer identification is a kind of behavioral biometrics using handwriting as the individual feature for personal authentication. Handwriting with a natural writing attitude is an effective way to represent the uniqueness of individual, and plays an essential role in the biometric traits [2].

Writer identification can be divided into text-dependent and text-independent approaches [3]. According to the common sense, the more characters are appeared in both query and reference, the better the performance is

© Springer Nature Singapore Pte Ltd. 2021
G. Zhai et al. (Eds.): IFTC 2020, CCIS 1390, pp. 194–205, 2021.
https://doi.org/10.1007/978-981-16-1194-0_17

achieved. When there are only few characters (e.g. phrase, name), the text content of the handwriting documents should be the same to ensure the credibility of writer identification. On the other hand, if there are lots of characters (e.g. textlines, paragraphs, or pages), the text content of handwriting documents could be the different. Compared to text-dependent approaches, it is well known that text-independent approaches have no limitations on character amount on the handwriting documents. Text-independent approaches require sufficient characters created by different people for features representation. If training samples are not enough, the approaches are not text-independent any more. Thus, the definition of text-independent is rather qualitative than quantitative. Actually the factors of text content are of crucial importance for the text-independent approachs, and influence the performance of an identification system to some extent.

Generally the researchers attempted to propose a novel feature which is distinctive and at the same time robust to changes under different conditions. Zhu et al. [4] extracted texture features with the two-dimensional Gabor filtering technique. In order to with reduce the calculational cost, Shen et al. [5] used the wavelet technique to improve the Gabor filters. Zhang et al. [6] proposed a hybrid method combining Gabor model with mesh fractal dimension. Li and Ding [7] focuses on different locations of linked pixel pairs, and presented the grid microstructure feature (GMF). Inspired by the idea of GMF, Xiong et al. [8] proposed a contour-directional feature.

In this paper, we try to investigate the text-independent approaches from another perspective. An empirical study of text factors and its effects on Chinese writer identification is presented. This study has two. The first one is to find the experimental minimum number of characters for reliable writer identification. Though the performance of identification not only depends on the number of characters, but also depends on the capacity of datasets and the quality of feature and handwriting, the minimum character amount is still quite useful and meaningful when we try to design a new dataset or test the performance of a new text-independent feature. The second goal is to make clear the relationship between the identification performance and text factors.

2 Experiments

Usually, researchers believe that the number of characters is the main factor which carry the greatest responsibility for good performance. This implies that with more characters in both query and reference, the performance will be better. Brink et al. [9] reported that the minimum character amount of needed text for text-independent writer identification is 100 characters (Dutch and English). In our opinion, text factors consist of two types of information: the number of characters in both query and reference (ACQR) and the number of the same

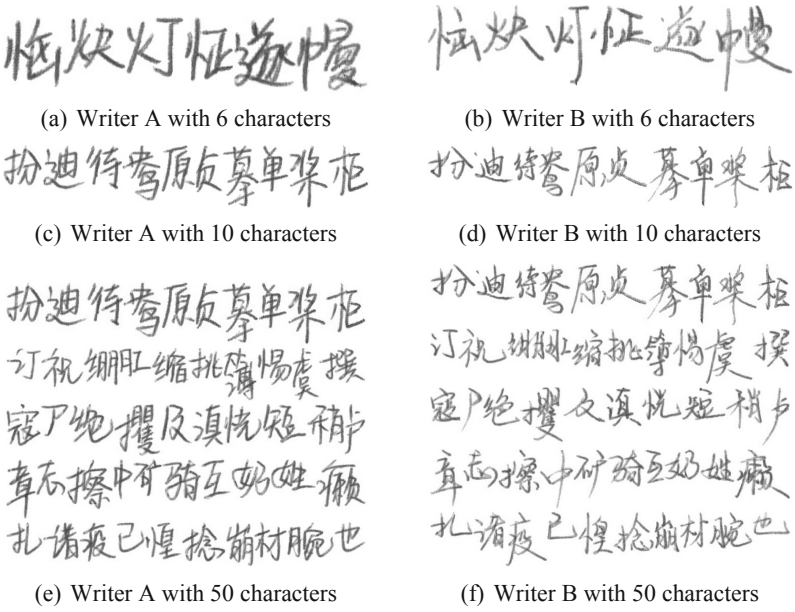
characters in both query and reference (ASCQR). They can reflect the integrated relationship of text content between the query and reference. We design two specific experiments to measure the impact of each factor on the performance and analyze the interactions between them.

2.1 Dataset

The HIT-MW dataset [10] is a famous Chinese dataset for writer identification. It includes 853 documents and 186,444 characters. Our experiments attempt to explore the impacts of text factors under different conditions, therefore the number of characters and text content in both query and reference should to be controlled accurately. The HIT-MW dataset cannot satisfy for this demand. In order to meet this requirement, the documents for our experiments are dynamic generated. We choose CASIA-OLHWDB1.0 [11] as the initial isolated character database to generate new document images which have alterable text content and characters. CASIA-HWDB1.0 contains isolated handwritten Chinese characters samples, and the samples were contributed by 420 persons, and each writer was asked to write 3,866 Chinese characters. 3,740 characters are in the GB2312-80 level-1 set which includes 3755 most frequently used Chinese characters. There are two obvious advantages for this dataset. Firstly, we have more than 1,680,000 characters from 420 writers to generate various documents. Only based on sufficient data, we can evaluate the impacts of each factor on writer identification correctly. Secondly, the character samples are isolated and each writer has the same 3,866 characters, thus we can have perfect control of text content in every documents to simulate the documents under different conditions.

2.2 Document Image Generation

A simple but effective method is employed to generate document images for our experiments. The generated document is line based handwriting document. In order to create a document of writer α with n characters $\{C_i | i = x_1, x_2, \cdots, x_n\}$, the isolated character samples $\{C_i^{\alpha} | i = x_1, x_2, \cdots, x_n\}$ are selected to create the text-lines. Each text-line $\{L_i | i = 1, 2, \cdots, \lceil n/10 \rceil\}$ is formed by a sequence of close connected characters. The generated document I is arranged by text-lines from top to bottom, and the space between adjacent text-lines are five pixels. Some generated images are shown in Fig. 1, and they are similar to the handwriting with a natural writing attitude.



(a) Writer A with 6 characters (b) Writer B with 6 characters

(c) Writer A with 10 characters (d) Writer B with 10 characters

(e) Writer A with 50 characters (f) Writer B with 50 characters

Fig. 1. Generated images of two writers with different characters

2.3 Contour-Directional Feature

In our previous work, we proposed a histogram based feature called as contour-directional feature (CDF) [8] for text-independent Latin writer identification. Our proposed CDF utilizes directional information to describe the distribution of pixel pairs, and represents the the writing style for different writers. We compare the CDF with other common text-independent features. There are the contour-hinge feature (CHF) [12], the multi-scale contour-hinge (MCHF) [13], the GMF [7] and the SIFT based features (SDS+SOH) [14]. As done in [7], we conduct the test using the handwritings of 240 writers in the HIT-MW dataset. Table 1 shows the writer identification performance of different features.

Table 1. The performance of different features on the HIT-MW dataset

#	CHF	MCHF	GMF	SDS+SOH	CDF
Soft-Top-1	84.6%	92.5%	95.0%	95.4%	**95.8%**
Soft-Top-5	95.4%	97.1%	98.3%	98.8%	**99.2%**
Soft-Top-10	96.7%	97.5%	98.8%	**99.2%**	**99.2%**

2.4 Evaluation Criterion

Conventional evaluation criterions for writer identification are the soft and hard TOP-N criterions. These criterions are appropriate and reflect the performance of the feature properly when the dataset is consisted of constant documents and the experiments only executed a few times. Our experiments are performed by repeatedly on the dynamic generated documents that make the traditional criterions cannot show all details of the results. Thus, we need to find another way to evaluate the performance. It is clear that the performance of identification is relevant to the separability of the feature, and the separability is represented by the inter-class distance and inner-class distance. For this reason, we decide to utilize the similarity distance of the features to represent the performance. Table 2 is a simulation comparison of different methods. *Method 1* has the best soft Top-5 accuracy, while *Method 2* has the best soft Top-10 accuracy and *method 3* has the best soft Top-1 accuracy. It is hard to determine which condition is better. If we concern about the inter-class and inner-class distance, the result is pretty obvious. The *Method 2* has larger inter-class distance and smaller inner-class distance, thus it maybe has batter performance. In other words, the similarity distance can be regard as the synthetic score of TOP-N criterions.

Table 2. A simulation comparison of different methods

#	*Method 1*	*Method 2*	*Method 3*
Soft-Top-1	68.2%	69.9%	**70.2%**
Soft-Top-5	**83.5%**	81.8%	82.4%
Soft-Top-10	86.8%	**89.4%**	88.6%
Inter-class distance	4.2E−06	5.5E−06	4.5E−06
Inner-class distance	2.4E−06	1.8E−06	2.2E−06

2.5 Experimental Setup

Text factors consist of two types of information: the number of characters in both query and reference (ACQR) and the number of the same characters in both query and reference (ASCQR). We design two experiments to analyze the relationship among the performance and ACQR and ASCQR. Experiment 1 is to calculate the similarity distance (SD) of the features extracted from the dynamic generated images with the same number of characters. There are three conditions between the query and reference: 1. They are from the same person and have totally different characters (SPDC); 2. They are from different persons and have the same characters (DPSC); 3. They are from the different persons and have totally different characters (DPDC). The SDs of SPDC, DPSC, and DPDC are computed by 126,000 ($420 * 300$) times repeated trials. It is noted

that the characters in each trial are randomly assigned, and the writer of the query in each trial of DPSC and DPDC is also randomly assigned.

When the number of characters in the reference is constant, the number of the same characters in both query and reference (ASCQR) also has an effect on the performance. Experiment 2 is to calculate the SD of the features extracted from the dynamic generated images with different number of the same characters. There are two additional conditions between the query and reference: 1. The numbers of the characters in the query (ACQ) is equal to the number of characters of the characters in the reference (ACR); 2. ACQ is not equal to ACR. The SDs of SPDC, DPSC, and DPDC in two conditions are also computed by 126,000 (420*300) times repeated trials. Through the statistical analysis of the result, we can find the relationship between the performance and ASCQR.

2.6 Experimental Results

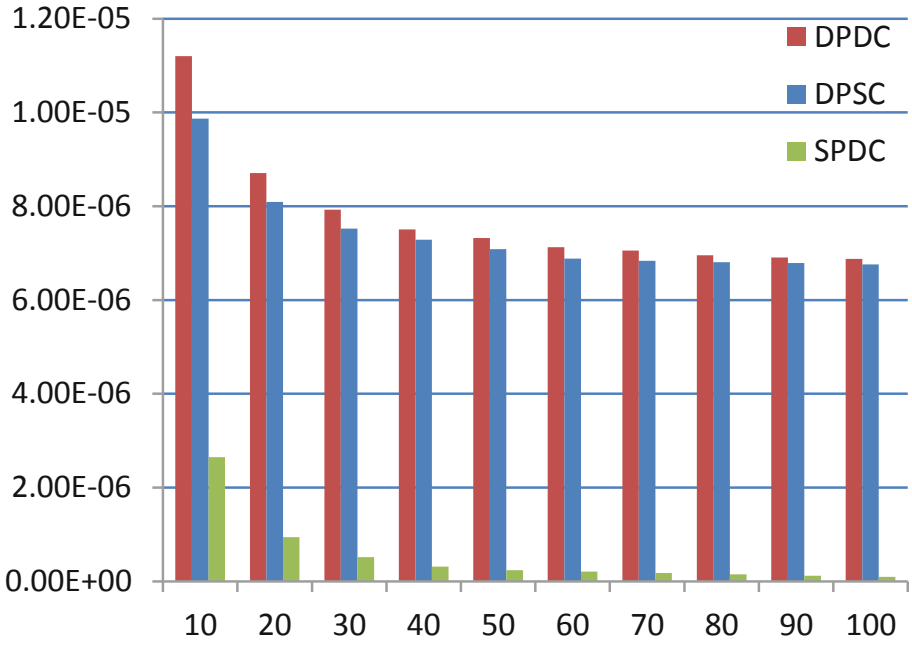


Fig. 2. Relationship between ACQR and SD

Figure 2 shows the statistical result of the Experiment 1 according to ACQR. The x-axis represents ACQR and the y-axis represents the mean value of the SD. As shown in figure, the SDs of all conditions become smaller when ACQR is increased. However, the decline rate of SPDC is much greater than the decline rates of DPSC and DPDC. This implies the separability of the feature becomes

better with increasing characters. On the other hand, the SD of DPDC is always larger than the SD of DPSC. It shows that the same characters in both query and reference can influence the similarity of the features. The difference of the SDs of DPDC and DPSC is smaller when ACQR is increased, and the SD of DPDC is approximately equal to the SD of DPSC when ACQR is larger than 50. It demonstrates the feature is really text-independent when ACQR is larger than 50.

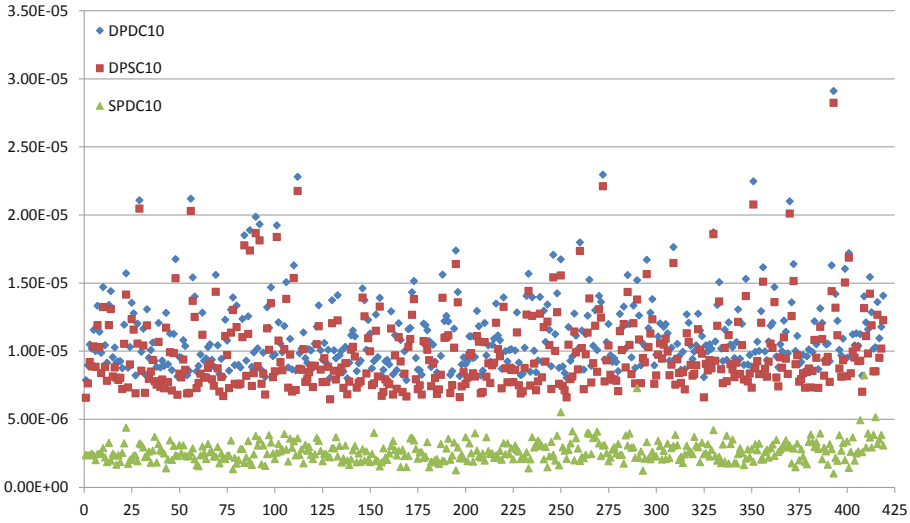


Fig. 3. Relationship between the writer and SD (ACQR = 10)

Above results are based on mean value of all 420 writers, they cannot reflect the difference of various writers. Thus, we do statistics on the result of Experiment 1 according to different writers. Figure 3, Fig. 4 and Fig. 5 show relationships between the writer and the SD when the ACQR is equal to 10, 50, 100. The x-axis represents the number ID of writers, and the y-axis represents the mean value of the SD. When ACQR is equal to 10, the magnitude of DPDC and SPDC is the same and most categories are separable. But there are still two categories whose SD of SPDC is larger than the SD of DPDC of some other writers. When ACQR is equal to 50 and 100, the SD of DPDC is approximately equal to the SD of DPSC, and the magnitude of DPDC is larger ten times than the magnitude of SPDC. It supports that separability the feature is text-independent when ACQR is larger than 50.

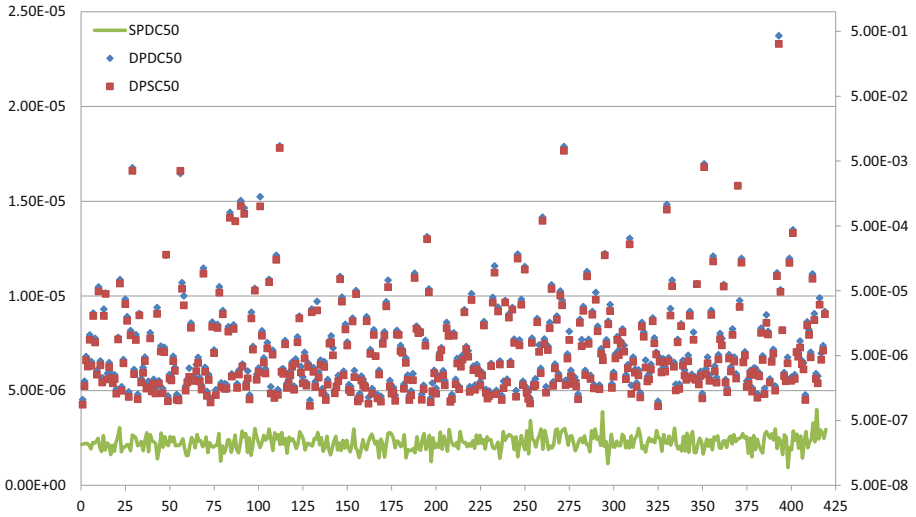


Fig. 4. Relationship between the writer and SD (ACQR = 50)

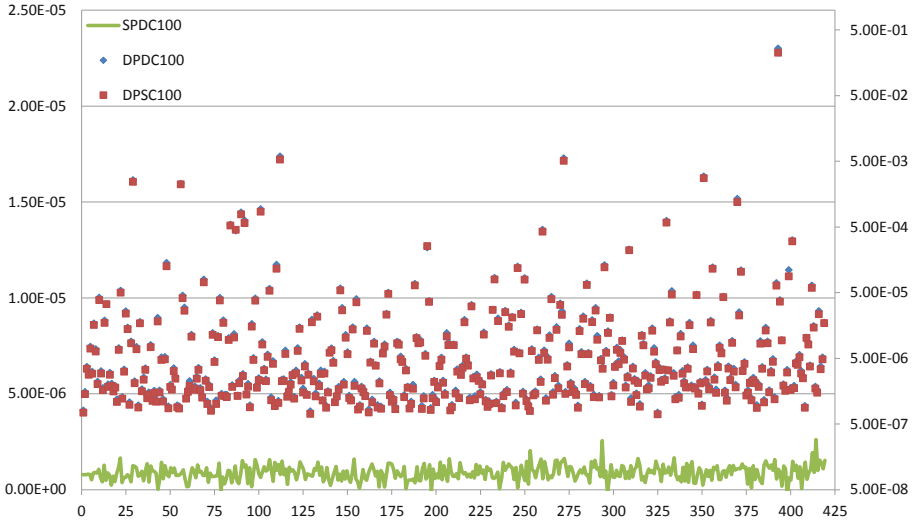


Fig. 5. Relationship between the writer and SD (ACQR = 100)

We also obtain the probability distributions of SD with different characters in both query and reference. Figure 6, Fig. 7 and Fig. 8 are the probability distributions of SD when the ACQR is equal to 10, 50, 100. The x-axis represents the value of the SD and the y-axis represents the probability of the occurrence. Figure 6 shows that the probability distribution of SPDC and DPDC are overlapped. However, there is no overlap between the probability distributions of

SPDC and DPSC in Fig. 7 and Fig. 8. It implies that if a fixed threshold is used to identify different writers, the classification accuracy is 100% when ACQR is larger than 50.

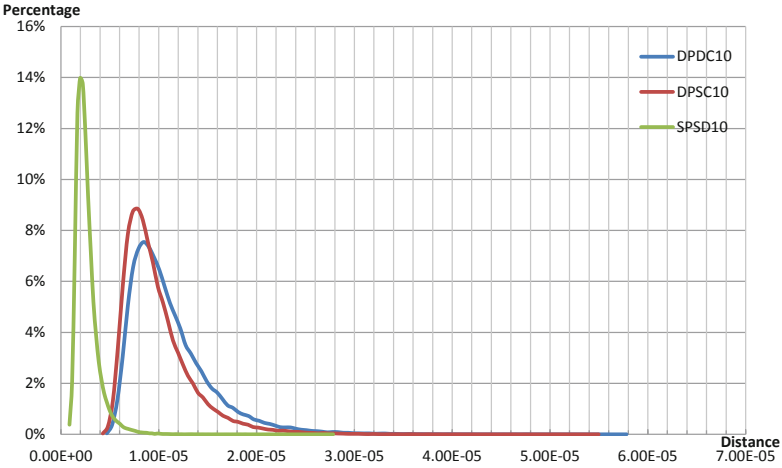


Fig. 6. The probability distribution of SD (ACQR = 10)

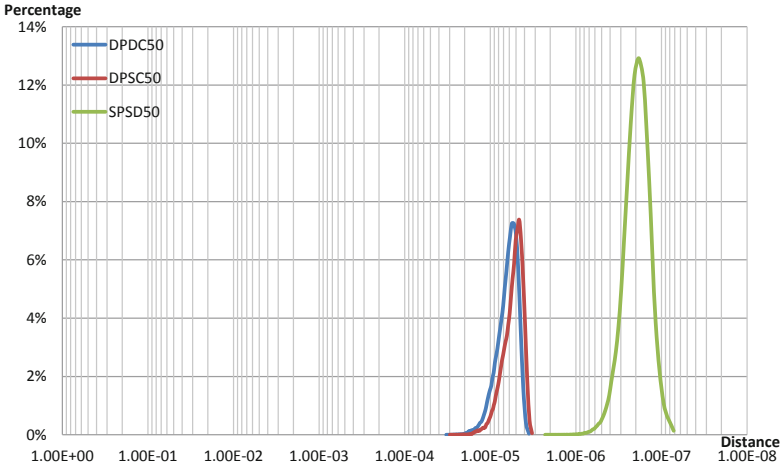


Fig. 7. The probability distribution of SD (ACQR = 50)

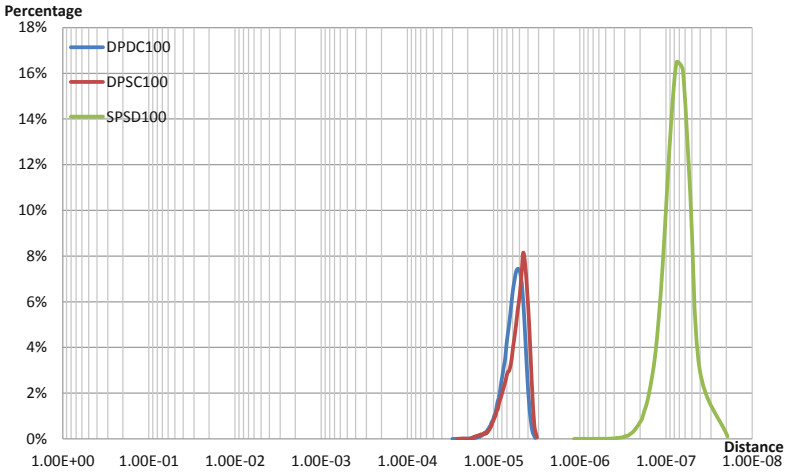


Fig. 8. The probability distribution of SD (ACQR = 100)

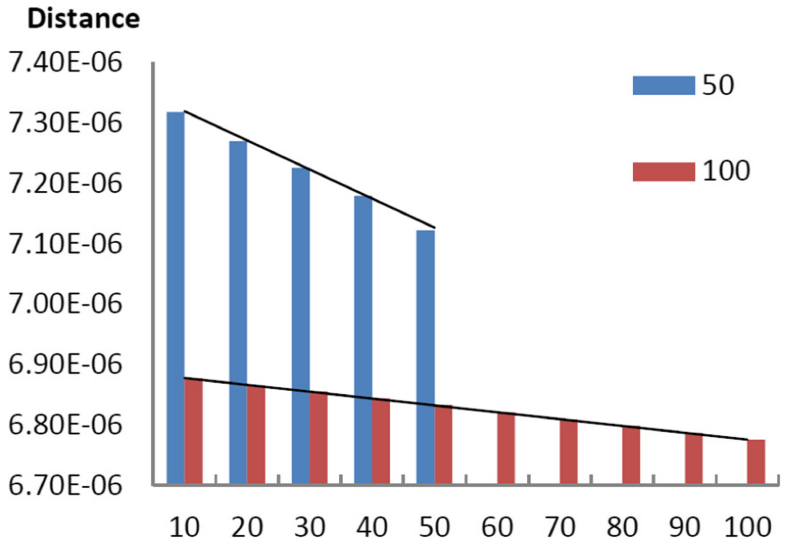


Fig. 9. The SD with different ASCQR(ACQ = ACR)

Figure 9 shows that the SD of the query and reference from different persons with different ASCQR when ACQ is equal to ACR, and the ACQR is 50 and 100. The x-axis represents the value of ASCQR and the y-axis represents the mean value of the SD. It is clear that the SD of different persons is reduced when ASCQR is increased. The relationship between the SD and ASCQR is linear, and the decline rate decreases with the increasing ACQR. Figure 10 shows that the SD of the query and reference from different persons with different ACR

when ASCQR is from 10 to 50. The x-axis represents the value of ACR and the y-axis represents the mean value of the SD. The figure indicates that the SD of different persons is reduced when ACR is increased and ASCQR is fixed. The decline rate of the SD decreases with the increasing ACR, and the decline rate of the SDs with different ASCQR and the same ACR seems to the same.

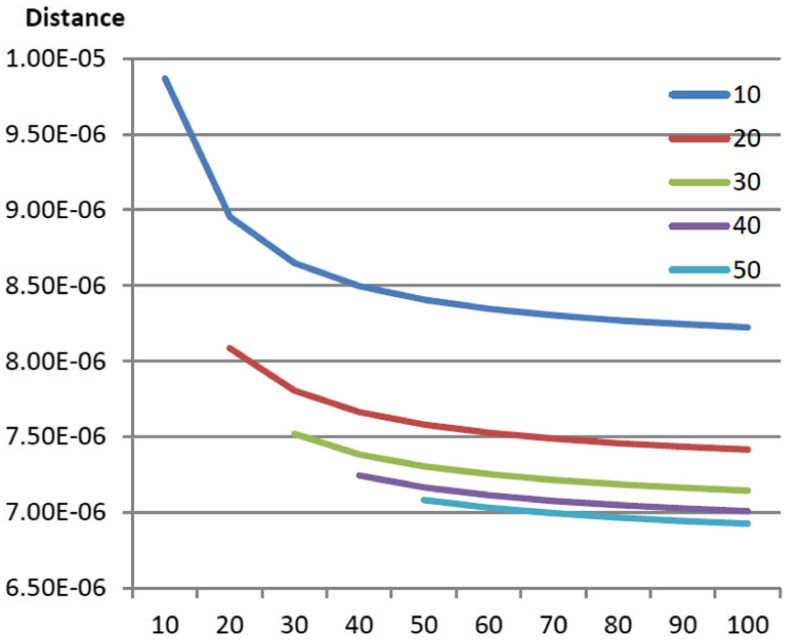


Fig. 10. The SD with different ACR(ACQ $\neq$ ACR)

3 Conclusion

As a rule of thumb, 50 characters is the minimum character amount for Chinese writer identification when using a strong text-independent feature such as CDF. In general, the more difficult the identification task, the more text are needed. More characters in handwriting documents are always better in every cases, even the amount of characters is only increased in the reference handwriting documents.

Acknowledgements. This work is jointly sponsored by National Natural Science Foundation of China (Grant No. 62006150), Shanghai Young Science and Technology Talents Sailing Program (Grant No. 19YF1418400), Shanghai Key Laboratory of Multidimensional Information Processing (Grant No. 2020MIP001), and the Fundamental Research Funds for the Central Universities.

References

1. Jain, A.K., Ross, A., Pankanti, S.: Biometrics: a tool for information security. IEEE Trans. Inf. Forensics Secur. **1**(2), 125–143 (2006)
2. Bulacu, M.L.: Statistical pattern recognition for automatic writer identification and verification. Ph.D. thesis, University of Groningen (2007)
3. Plamondon, R., Lorette, G.: Automatic signature verification and writer identification - the state of the art. Pattern Recogn. **22**(2), 107–131 (1989)
4. Zhu, Y., Tan, T.N., Wang, Y.H.: Biometric personal identification based on handwriting. In: Proceedings of the International Conference on Pattern Recognition, pp. 797–800 (2000)
5. Shen, C., Ruan, X.G., Mao, T.L.: Writer identification using Gabor wavelet. In: Proceedings of the World Congress on Intelligent Control and Automation, pp. 2061–2064 (2002)
6. Zhang, J., He, Z., Cheung, Y., You, X.: Writer identification using a hybrid method combining gabor wavelet and mesh fractal dimension. In: Corchado, E., Yin, H. (eds.) IDEAL 2009. LNCS, vol. 5788, pp. 535–542. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-04394-9_65
7. Li, X., Ding, X.: Writer identification of Chinese handwriting using grid microstructure feature. In: Tistarelli, M., Nixon, M.S. (eds.) ICB 2009. LNCS, vol. 5558, pp. 1230–1239. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-01793-3_124
8. Xiong, Y.-J., Wen, Y., Wang, P.S., Lu, Y.: Text-independent writer identification using sift descriptor and contour-directional feature. In: Proceedings of the International Conference on Document Analysis and Recognition, pp. 91–95 (2015)
9. Brink, A., Bulacu, M.L., Schomaker, L.: How much handwritten text is needed for text-independent writer verification and identification. In: Proceedings of the International Conference on Pattern Recognition, pp. 1–4 (2008)
10. Su, T.H., Zhang, T.W., Guan, D.J.: Corpus-based HIT-MW database for offline recognition of general purpose Chinese handwritten text. Int. J. Doc. Anal. Recogn. **10**(1), 27–38 (2007)
11. Liu, C.-L., Yin, F., Wang, D.-H., Wang, Q.-F.: Online and offline handwritten Chinese character recognition: benchmarking on new databases. Pattern Recogn. **46**(1), 155–162 (2013)
12. Bulacu, M.L., Schomaker, L., Vuurpijl, L.: Writer identification using edge-based directional features. In: Proceedings of the International Conference on Document Analysis and Recognition, pp. 937–941 (2003)
13. Van Der Maaten, L., Postma, E.: Improving automatic writer identification. In: Proceedings of the Belgium-Netherlands Conference on Artificial Intelligence, pp. 260–266 (2005)
14. Wu, X.Q., Tang, Y.B., Bu, W.: Offline text-independent writer identification based on scale invariant feature transform. IEEE Trans. Inf. Forensics Secur. **9**(3), 526–536 (2014)