Research on online signature verification based on anomaly detection

Xinghua Xia, Shuang Wang

School of Information and Control Engineering, Shenyang Jianzhu University, Shenyang City, Liaoning Province, 110168, China.

Corresponding author’s e-mail: xxh8787@sjzu.edu.cn

Abstract: With the advancement of modern information technology, personal information security has become the focus of social attention. As a biological behavior feature, online signatures have been trained for a long time, have very personal characteristics, and are collected as a real-time sequence. In view of this characteristic, we can know that it is difficult for forgers to imitate the intrinsic and essential characteristics of the signer. In addition, a signature is a rapid writing process driven by the central nervous system and personal writing habits. Even the signer himself may write an unstable and similar forgery signature, so it is necessary to analyze the variability and stability of signatures. Therefore, this paper proposes a method of using the variance and characteristics of local features to find a stable segment of the user’s signature for authentication. In addition, we also adopt a two-stage authentication strategy that combines the above method with the traditional anomaly detection algorithm. An experiment was conducted on the open-access online signature database MCYT, and an EER of 2.63% was obtained. The experiment proved the effectiveness and robustness of the method.

1. Introduction

The rapid development of information technology brings great convenience to people's daily life, but also brings unprecedented challenges to network security. Therefore, real-time and accurate personal identity authentication is particularly important. Traditional identity attestation is based on the characteristics that oneself hold such as password, ID card, IC card to attestation, but the password may be stolen, forgotten, IC card may be lost, stolen, thus security is not high, in the use of users already feel inconvenient. The identity authentication based on human biometrics has been applied more and more because it can overcome the above shortcomings. Based on this background, a new technical field was established-called biometrics. Online signature verification is a "one-to-one" match between a test sample and a specific target, which solves the problem of "Is it someone".

As a feature of biological behavior, when the signature is collecting data, due to changes in the internal and external environment, even under multiple input by the same user, the external form of the signature and various real-time parameters of the signature will change to varying degrees. Moreover, with the passage of time, with continuous learning and changes in writing habits, signatures will not maintain high consistency for a long time, which increases the difficulty of online signature verification; second, if the forger is familiar with The various habits of the signer in the signing process, such as the time of signing, the severity of pen writing, etc., then this type of forgery is called skilled forgery. Skilled forged signatures have a high degree of similarity with real signatures, which is the biggest obstacle to signature verification.
2. Related work
Since the development of online handwritten signatures, many scholars such as Plamondon R and Impedovo have tried different methods to improve the performance of the system and achieved fruitful results: in 1964, Murray Eden[1] introduced a handwritten signature model, which was applied to a word. In the recognition system, although the model is not used to deal with the issue of signature verification, he emphasizes the need to observe the dynamic process of writing-reflecting the essence of "online recognition". Dynamic Time Warping (DTW) has become the most commonly used matching method in the signature verification system because it solves the similarity calculation problem when the timing lengths of two samples do not match. The team from Turkey used DTW-based principal component analysis (PCA) to obtain the best EERs of 2.84% and 2.89%. Fierrez [2] designed a 7-dimensional feature descriptor for each sampling point, using the HMM system to effectively improve the accuracy of the system, and finally obtained an equal error rate of 6.9% on the SVC2004 database.

Research on online signature verification by domestic researchers began in the 1990s. Cai Hongbin of the University of Science and Technology of China and others used wavelet transform to extract the inflection points in the position coordinate curve as the feature quantity, and used DTW to match the features, and obtained the recognition result of FRR=9 and FAR=0; Zhang Kui of Huazhong University of Science and Technology and others for simplicity in the process of information processing of dynamic features, a method of extracting special point sequences from characteristic curves as feature vectors is proposed. In addition, Qiu Yiming proposed an online signature verification method based on curve similarity. The author regards online signature trajectory data as a plane curve, and gives a curve similarity calculation model framework from curve similarity definition, similarity transformation and similarity distance.

3. Online signature verification based on anomaly detection
3.1 Preliminary preparation
3.1.1. Pretreatment. Preprocessing is to eliminate various interferences and distortions generated in the data collection process, and to standardize the position of the signature and the number of sample points. In this paper, the five-point cubic smoothing method will be used to smooth the signature data, and the cubic spline interpolation method will be used to resample the signature.
3.1.2. Feature extraction. Features are generally divided into two categories: global features and local features. The former describes the characteristics of the signature as a whole and globally, such as signature aspect ratio, average speed, signature time, etc. The latter describes the dynamic process of signatures, which are some time series functions generated during the signature formation process. This paper subjectively extracted 6 local features and 37 global features as the original feature set: \( F_l = \{ f_1, f_2, \ldots, f_6 \} \) and \( F_g = \{ f_7, f_8, \ldots, f_{43} \} \), where, \( f_1 \) : X-direction displacement, \( f_2 \) : Y-direction displacement, \( f_3 \) : pen pressure, \( f_4 \) : X-direction velocity, \( f_5 \) : pen speed, \( f_6 \) : path tangent angle. Due to the length of the article, other features are no longer described in detail.

3.2 Method description
Time series anomaly detection methods mainly include the following: 1) Model-based detection method: establish a data model, anomalies are those objects that cannot fit the model perfectly; 2) Cluster-based...
detection method: use directly or indirectly Existing clustering algorithms such as DBSCAN, K-means, ROCK, KNN [3] cluster the data, and treat those classes with few data points or data that cannot be clustered as anomalies; 3) Density-based detection Method: Calculate the density estimation of objects directly, especially when there is a measure of proximity between objects. Objects in low-density areas are relatively far away from their neighbors and may be regarded as anomalies [4].

3.2.1. Stage 1: one-class SVM. The idea of one-class SVM is [5] to find a hyperplane where all or most of the target samples are on one side, but not the target samples on the other side. One-class SVM also draws on the principle of SVM, using the kernel function to map the original low-dimensional feature space \( R^n \) to the high-dimensional space \( H \), Namely: \( x_i \rightarrow \phi(x_i) \). Through the target sample training, a hyperplane represented by the support vector can be obtained, which can separate the target sample from the origin as much as possible, and maximize the distance from the origin to the hyperplane under the condition of correctly distinguishing the target samples as much as possible. Define \( w \cdot \phi(x_i) - b = 0 \) as a hyperplane in a high-dimensional space, then the above process is transformed to solve the following optimization problems:

\[
\begin{align*}
\min F(w,b,\xi) &= \frac{1}{2}\|w\|^2 + \frac{1}{\nu N}\sum_{i=1}^{N}\xi_i - b; \text{s.t. } (w \cdot \phi(x_i)) \geq b - \xi_i, \forall i, \xi_i \geq 0
\end{align*}
\]

Among them, \( w \in H, b \in R \) is the hyperplane parameter, \( \nu \in (0,1] \) is the compromise between the number of misclassified samples and the distance from the origin to the hyperplane, and \( \xi_i \) is the slack variable.

3.2.2. Stage 2: Use the variance of local features to find the stable region of the signature. In statistics, the variance represents the degree of dispersion of the sample. The smaller the variance, the lower the degree of dispersion of the sample, which means that the sample data is more stable. Using this characteristic of variance, select a certain number of sample signatures, extract local features from them, and calculate the variance of their features, set a certain variance threshold Thres and specify: if \( \text{var} < \text{THRES} \), then \( \text{var} \in \text{stab_set} \), otherwise \( \text{var} \in \text{stab_lea} \). The stable point sets \( \text{stab_set} \) of the local features of multiple authentic and false signatures intersect to obtain \( \text{Arrset_true} \) and \( \text{Arrset_forg} \). Considering the influence of the forged signature in this segment, the stable area of the user is shown in equation (2).

\[
\text{SET} = \text{Arrset_true} - \text{Arrset_forg}
\]

Set the weights \( w_1, w_2 \), between the stable pen segment and the unstable pen segment, and calculate the similarity. In addition to the signature \( f_1, f_2 \), features to calculate the Euclidean distance, the four features of \( f_3, f_4, f_5, f_6 \) have no spatial geometric meaning in calculating the Euclidean distance, so the method of calculating the absolute value of the distance is adopted for authentication.

\[
\begin{align*}
\text{abs}_{\text{cmp}_{S_j}(x_i)} &= \text{temp}_{S_j_{-}f_i} - \text{temp}_{S_j_{-}f_j} \\
\text{Dist}_{\text{cmp}_{S_j}(x_i)} &= \frac{1}{N}\sum_{i\in \text{SET}}(w_i \text{dist}_{\text{sta}} + w_i \text{dist}_{\text{lea}})
\end{align*}
\]

\( \text{dist}_{\text{sta}} \) and \( \text{dist}_{\text{lea}} \) respectively represent the Euclidean distance between the sampling points in the stable pen segment and the unstable pen segment.

This forms a threshold matrix \( \text{THRES} \) with a dimension of \( 5 \times 5 \), where \( \mu_i \) and \( \sigma_i \) represent the mean and standard deviation of the distance between each template signature and other template signatures, and \( k \) is the substitution parameter.

\[
\text{th}_i = \mu_i \pm k \times \sigma_i \quad i,j = 1, \ldots, 5
\]

When inputting the signature to be tested, use equation (3-5) to calculate the distance matrix \( \text{Test}_{\text{sig}_{ss}} \) of the test signature, and perform the difference operation with \( \text{THRES} \). The difference is greater than 0 and recorded as 1, otherwise it is recorded as 0, and a number of 1s and 0s are obtained.
The composed judgment matrix \( \mathbf{f} \) satisfies equation (6) and considers the signature to be tested as a real signature, otherwise it is a forged signature. \( R \) and \( C \) indicate the number of rows and columns.

\[
\text{true} \_ \text{sig} = \begin{cases} \text{fea} \_ \text{sig} \forall 1 \geq 16 \cup R \leq 1 \cup 4 \cup C \leq 1 \cup 4 \end{cases} \tag{6}
\]

4. Experiment and result analysis

All experiments in this paper are carried out on the open access database - Sig_MCYT[6].

Stage 1: select 5 real signatures for each user, 7 forged signatures as the training set, and extract 37 global features for model training of the SVM classifier, and select the linear kernel function to use the remaining 38 signatures as the test set for experimentation. The authentication accuracy rate is shown in Figure 1.

![Figure 1. SVM certification accuracy rate.](image)

Stage 2: select 10 authentic and false signatures of each user in MCYT, extract 6 local features in \( \mathbf{F} \), and calculate their variances respectively to obtain a stable region \( \text{SET} \). Figure 2 shows that each user's \( \text{SET} \) is in each 3 Results marked on a real signature.

![Figure 2. Marking different users.](image)

After getting \( \text{SET} \), input a certain signature to be tested and set parameter \( w_1 = 0.3 \), \( w_2 = 0.7 \), \( k_1 = -1.25 \), \( k_2 = -1 \), \( k_3 = 0.25 \), \( k_4 = -1.5 \), \( k_5 = 0.75 \). The false rejection rate (FRR), false recognition rate (FAR) and equal error rate (EER) of each user are shown in Figure 3.

![Figure 3. (a) is the false rejection rate, (b) the false recognition rate, (c) and other error rates.](image)

Through the above experimental results, it can be found that a single method is difficult to achieve a higher authentication accuracy rate. The essence of machine learning is to make the local optimal, so the two methods are finally combined, that is, first use the classic OC-SVM to experiment, observe the authentication result of each user, and use the variance and variance of the local characteristics of the
authentication error signature. The method of finding the stable area of the signature is verified again, and the experimental results are shown in Figure 4. Table 1 shows the EER at each stage.

| Only stage 1 | Only stage 2 | Two-stage certification |
|--------------|--------------|-------------------------|
| EER          | 0.2162       | 0.1768                  |

Table 1. Experimental comparison results.

![Two-stage authentication error rate.](image)

Table 2. Compare the proposed method with the method in recent years.

| Database | Method | EER(%) |
|----------|--------|--------|
| SVC2004  | DTW    | 2.26   |
| MCYT     | Histogram-based features and Manhattan distance\[7\] | 4.02   |
| SUSIG    | Histogram-based features and Manhattan distance | 2.94   |
| MCYT     | Proposed method(SVM+ based on local feature variance and its Characteristics) | 2.63   |

Through the comparison of some experiments listed in Table 2, it can be found that the average equal error rate of 100 users can basically meet the requirements. However, observing Figure 4 can find that the error rate of users 2, 8, 41, and 47 is still above 10%. This shows two problems: first of all, for users whose authentication accuracy is not high, it shows that the "stable" we found is not very accurate. The second is that the features we selected are not representative of this part of users, which leads to low accuracy. Next, we will focus on researching the signature data of these users or selecting other features or changing methods until all forged signatures are found.

5. Conclusion

In statistics, the variance represents the degree of dispersion of the sample. In view of this characteristic, in this article, we propose a way to use the variance of local features and their characteristics to find the stable area of the signature, and set the stable part and other parts Different weights are used to determine the algorithm of the authentication threshold. Finding those signatures that do not meet the threshold is a forged signature. In order to reduce external interference, before authentication, we have carried out effective pre-processing on the signature, including smoothing, correcting and resampling. Then two more consistent features are extracted. In the authentication link, the two features were used to conduct experiments respectively. We found that a single algorithm could not be applicable to all users, so this paper adopted a two-stage authentication strategy, that is, in the first stage, the traditional global feature authentication method, such as SVM, was used for authentication. Secondly, the local special authentication method is used to authenticate the wrong signatures in the first stage for the second stage. The experiment is also carried out on MCYT database, and the average error rate of 2.63% is obtained.

Signature authentication technology is a frontier topic in pattern recognition, artificial intelligence and other fields, which has aroused the research interest of many scholars at home and abroad. At present, it has become a convenient and efficient personal identity authentication technology, and has a broad application prospect in government departments, finance, transportation and other fields. Compared
with other biometric identification methods, signature authentication has considerable advantages, such as good uniqueness, non-aggression, easy to be accepted, etc., but its main disadvantage is the instability of the signature itself. Therefore, efforts should be made to overcome this disadvantage in the future research, which is expected to further improve the accuracy of authentication.

References

[1] Merlmeister, P., Eden, M. (1964) Experiments on computer recognition of connected handwritten words. Information and Control, 7(2).
[2] Fierrez, J., Ortega-Garcia, J., Ramos, D., Gonzalez-Rodriguez, J. (2007) HMM-based on-line signature verification: Feature extraction and signature modeling. Pattern Recognition Letters, 28(16).
[3] Zhu, L., Qiu, YY., Yu, S., Yuan, S., (2017) A fast KNN-based MST outlier detection method. Chinese Journal of Computer, 40(12): 224-238.
[4] TAN, B., HE, H. (2017) A local density-based approach for outlier detection. Neurocomputing, 241(7): 171-180.
[5] YE J. Research on online signature authentication based on support vector machine and single-class support vector machine [D].South China University of Technology,2016.
[6] Diaz, M., Miguel, A., Donato, F., Muhammad Imran Malik, I., Pirlo, G., Plamondon, R. (2019) A Perspective Analysis of Handwritten Signature Technology. ACM Computing Surveys (CSUR), 51(6).
[7] Sae-Bae, N., Memon, N.(2014) Online signature verification on mobile devices. IEEE Trans. Inf. Forensics Security, vol. 9, no. 6, pp. 933–947.