Improved Keystroke Authentication Accuracy Based on Statistics and Weight

Li Jian, Guo Xiaojing, Li Meiyun, Li Ruifan
School of Computer, Beijing University of Posts and Telecommunications, Beijing 100876, P. R. China

Abstract: In order to improve the recognition accuracy of keystroke authentication, a methodology based on feature extraction of keystroke sequence is presented in this paper. Firstly, the data of the users´ keystroke feature information that has too much deviation with the mean deviation is filtered out. Secondly, the probability of each input key is calculated and 10 values which do not have the best features are selected. Thirdly, they are weighed and a score evaluating the extent to which the user could be authenticated successfully is calculated. The benefit of using a third-party data set is more objective and comparable. At last, the experimental result shows that the detector in this paper is more accurate than the other three.

Key words: authentication; biometrics; keystroke dynamics

I. INTRODUCTION

At present, many application systems and softwares are still using a static password authentication technology in the login process because of its low cost and convenience. However, this authentication is fragile when there is a careless user or a weak password. The purpose of this paper is to improve the login-password authentication recognition accuracy using biometric characteristic. The advantage is that biometric characteristics are unique to each person and could not be stolen, lost or forgotten.

In earlier keystroke dynamics-based systems, many classifiers are used to achieve higher system accuracy. Statistics [1-3], neural networks [4], fuzzy logic [5], support vector machines [6], k-nearest neighbor [7], and other classifiers [8-17] have been developed into Keystroke Dynamics-Based Authentication (KDA) systems. In 2008, Kang et al. [3] used artificial rhythms to improve the quality of keystroke data. However, users in the system have to remember the locations of the inserting pauses in their password, which is not an innate typing characteristic. In addition, Wahyudi Martono et al. [6] discussed the design and development of keystroke pressure-based typing biometrics for individual user's verification, but this method requires specially modified keyboards which may not be widely acceptable, as these keyboards may not be common in practical use. Recently, Saurabh Singh and Dr.K.V.Arya [7] proposed a new approach in free text keystroke authentication based on key classification, and get an encouraging result; however, this method is not compatible with our password authentication and other fixed text, such as on-line banking system.

Detector in this paper is named as TOP10 and comparisons between our method and the other three proposed in Refs. [18-20] are made. Firstly, comparisons on anomaly detection detector for keystroke dynamics were made by Kevin S. Killourhy and Roy A.Maxion [18], which includes 14 keystroke authentication detectors [3, 21-22]. As a consequence, the performance of a detector named Man_Scaled outperforms all the other detectors, and its Equal Error Rate (EER) is 0.096. Moreover, Yang Haiyan et al. [19] proposed a methodology called Seq_Dist. It adopts a new representation for the duration of tri-char for the user's input feature, using the relative distance of user's keystroke speed standing for its relative rhythm. At last, they reach a recognition accuracy whose False Acceptation Rate (FAR) is 2.457% and False Rejection Rate (FRR) is 0.05%. The Last detector is called Diff_Sub, which is based on difference subspace in Ref. [20]. It calculates the common feature vector to depict a user's keystroke pattern according to keystroke features of the user's several recent successful authentications, and then utilizes the Euclid distances between current keystroke pattern vectors and common feature vectors to identify the users. As a result, it obtains a result with FAR = 1.3% and FRR = 2%.

II. PAGE LAYOUT

This paper is organized as follows. Section II describes our methodology to improve the recognition accuracy of keystroke authentication. In Section III, the experiments are presented and the results are discussed. Finally, conclusions and further work are discussed in Section IV.

III. PROCEDURE OF TOP10

3.1 Data collection

Keystroke dynamics, the analysis of typing rhythms to discriminate among users, is based on the hypothesis that different person would type in different typing rhythm. In Ref. [21], these significant patterns of different people typing in the same phrase of text can be clearly visualized. In fact, the difference is mainly manifested in keystroke duration and keystroke latency. Besides, Ref. [22] has proved that utilizing two factors would be more effective than only using one of two, so both time latency and time duration are extracted as keystroke feature in the paper.

The password of n bits contains $(2n-1)$ keystroke eigenvalues, n of which is the number of keystroke latency, and the other is the number of keystroke intervals. Meanwhile, they also constitute a vector of $(2n-1)$-dimensional feature space. According to Ref. [23], the benchmark data set was completed by 51 subjects (typists). Each subject typed the same password 400 times over 8 sessions (50 repetitions per session). They waited at least one day between sessions to capture some of the day-to-day variation of each subject's typing. The password (.tie5Roanl) was chosen to be the representative of a strong 10-character password.

3.2 The procedure

Step 1. The keystroke sequence is divided into two vectors $A_{mn}(pp)$ and $A_{mn}(pr)$, where $A_{mn}(pp)$ stands for the elapsed time of a key from down to up, while $A_{mn}(pr)$ means the interval of the down time of two successive keys, so this value may be negative in some cases.

Step 2. According to the hypothesis of Bleha Saleh [24] and Leggett [5], the user's input time data satisfies the normal distribution. So m vectors collected from a user's input are organized as a sample matrix $A_{mn} = [\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \cdots, \boldsymbol{\alpha}_m]^T$, in which $\boldsymbol{\alpha}_i$ is an n-dimensional row vector, then the mean value μ_k and standard deviation σ_k for each column can be calculated as follows,

$$\mu_k = \frac{1}{m}\sum_{i=1}^{m}\alpha_{ik};\ k \in (1,n) \tag{1}$$

$$\sigma_k = \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}(\alpha_{ik}-\mu_k)^2} \tag{2}$$

According to the meaning of the normal distribution curve, the area covered by the normal probability distribution is 0.997 4.

Then, $\Delta = \alpha_{ik} - \mu_k$ for each element α_{ik} is calculated. If $\Delta > 3\sigma_k$, then α_{ik} should be removed from the vector $\boldsymbol{\alpha}_i$ and replaced by μ_k. This process will be repeated until no element satisfies $\Delta > 3\sigma_k$. Finally μ_k and σ_k are denoted as μ and σ separately. This method could filter out those inputs which vary greatly with the usual keystroke for each user, namely, to achieve the purpose of reducing the possibility of misleading by those data.

Step 3. For the given vector $t = (t_1, t_2, \cdots, t_n)$, the probability of $t_k \in (t_1, t_2, \cdots, t_n)$ could be obtained by Eq. (3).

$$p(t_k) = 1 - 2\int_{\mu_k}^{\mu_k + |\mu_k - t_k|} \frac{1}{\sqrt{2\pi}\sigma_k}$$

$$\exp\left[-\frac{1}{2}\left(\frac{t_k-\mu_k}{\sigma_k}\right)^2\right]dt \qquad (3)$$

However, in the use of C++ programming by Microsoft Visual Studio 2005 development platform, it is more convenient to get the probability at a specific point in the test vector by the cumulative distribution function in C++ Boost library as follows,

$$p(t_k) = 2 \times \mathrm{CDF}(\mu_k - \mathrm{ABS}(\mu_k - t)) \qquad (4)$$

Using the method can avoid the complicated process of quadrature. In Eq. (4), ABS (absolute) means obtaining the absolute value, and Cumulative Distribution Function (CDF) calculates a variable´s probability that its value is less than or equal to variable value x, which is also the integration of the probability density function from $-\infty$ to the value x.

Step 4. We Select the 10 minimum probability values obtained from Step 3 and denote them as(s_1, s_2,⋯,s_{10}) , and then weigh them at Step 5.

Step 5. The procedure of weighing is as follows:

1) For a given vector(s_1,s_2,⋯,s_n) (set n=10), its weight factor should be set as $w=\left(\frac{1}{n},\frac{1}{n},\cdots,\frac{1}{n}\right)$ for their same occurrence probability in practical environment. So the mean value and standard deviation of the above vector can be calculated by Eq. (5) and Eq. (6) separately.

$$\bar{s} = \frac{1}{n}\sum_{i=1}^{n} s_i \qquad (5)$$

$$\sigma_s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\bar{s}-s_i)^2} \qquad (6)$$

Consequently, the mean $\bar{s}$ and standard deviation σ_s can determine a normal distribution $N(\bar{s},\sigma_s^2)$.

2) Get a new vector(p_1, p_2, ⋯, p_n) by making standardization utilizing the mean $\bar{s}$ and standard deviation σ_s in Eq. (7) for the given vector(s_1,s_2,⋯, s_n) .

$$p_i = \frac{1}{\sigma_s}(s_i - \bar{s}) \qquad (7)$$

3) Obtain the probability of $p_i \in (p_1, p_2, \cdots, p_n)$ in the normal distribution $N(\bar{s},\sigma_s^2)$ by Eq. (8), and name the probability vector as(q_1, q_2, ⋯, q_n) .

$$q(p_i\,|\,\bar{s},\sigma_s) = \frac{1}{\sqrt{2\pi}\,\sigma_s}\exp\left(-\frac{(p_i-\bar{s})^2}{2\sigma_s^2}\right) \qquad (8)$$

4) Using the above(q_1, q_2, ⋯, q_n) , get a weight vector(w_1, w_2, ⋯, w_n) as follows.

$$w_i = \frac{q_i}{\sum_{j=1}^{n} q_j} \qquad (9)$$

Step 6. In the end, it is easy to obtain the score evaluating the user by using Eq. (10). Whether the user could be authenticated successfully or not depends on the score s of her/his input password.

$$s = \frac{1}{n}\sum_{i=1}^{n} s_i w_i \qquad (10)$$

Ⅳ. COMPARING THE FOUR DETECTORS

4.1 Method of using the benchmark data set

During the process of comparing the four detectors, a benchmark data set for keystroke dynamics supported by Kevin Killourhy and Roy Maxion [23] is used, which is a supplement to Ref. [18]. Using a third-party data to compare the various detectors is more objective and the results will be more convincing.

It is started by designating one of our 51 subjects as the genuine user, and the rest as impostors. Next, recognizing the genuine user and impostors by training an anomaly detector and testing its ability as follows.

1) Firstly, the training phase of the detector on the timing vectors from the num= 5, 10, 15, 20, 25, 30, 35, 40 and 45 password repetitions typed by the genuine user are ran, where num denotes the number of samples used in experiments below. So the detector builds a model template of the user´s typing behavior.

2) Then, we conduct the test phase of the detector on the timing vectors from the remaining 45, 40, 35, 30, 25, 20, 15, 10 and 5 repetitions typed by the genuine user. Record the anomaly scores assigned to each timing vector as user scores, namely FRR. Taking 40 as an example, it will produce 8 × 40 = 320 records.

3) Finally, we conduct the test phase of the detector on the timing vectors from all the repetitions typed by the other 50 impostors. Then record the anomaly scores assigned to each timing vector as

impostor scores, namely FAR. The number of attacks record is the same for each user, and the value is 50 × 8 × 50 = 20 400.

Repeat this process and designate each of the other subjects as the genuine user in turn. Still taking 40 as an example, a totality of 16 320 sets of user (51 × 8 × 40) and 1 020 000 sets of impostor (51 × 50 × 8 × 50) could be obtained after the training and testing of each detector.

In a real login process for any authentication system, it is impractical to input his password as many as 200 times for a user in order to implement complex algorithms in building a template. As a result, the template is generated by trying to increase the user's input times of password of various tested detectors.

4.2 Comparison result

All the scores of each detector are imported into database, and then FRR and FAR are calculated by Eq. (11) and Eq. (12) separately, in which A denotes the acceptance times of impostor, D means denial times of genuine user, and T is for total repetitions of genuine user.

$$FRR = \frac{D}{T} \tag{11}$$

$$FAR = \frac{A}{T} \tag{12}$$

Ideally, the values of FRR and FAR should be as low as possible. However, reducing the FRR value is supposed to increase the success chances of the attacker, while blindly decreasing FAR will lead to higher denial of legitimate users because of strict identification standards in turn. Therefore, both FRR and FAR should be adjusted according to actual requirements. In addition, EER is a point where FRR equals to FAR, so the lower the EER is, the more accurate the detector is.

Also taking num= 45 as an example, the values of EER, FAR, FRR, FAR + FRR of four detectors are shown in Table I.

The following four graphs show the values of EER, FAR, FRR and FAR + FRR of each detector when the num= 5, 10, 15, 20, 25, 30, 35, 40 and 45. Moreover, as for the fitting curve of FAR and FRR of Seq_Dist, there is no discrete point at the intersection, so its EER does not exist in Table I and Figure 3.

From the four graphs in Figures 1 ~ 4, we can see that the authentication accuracy of each detector continues to improve with the increase of the number of samples, especially TOP10, which has

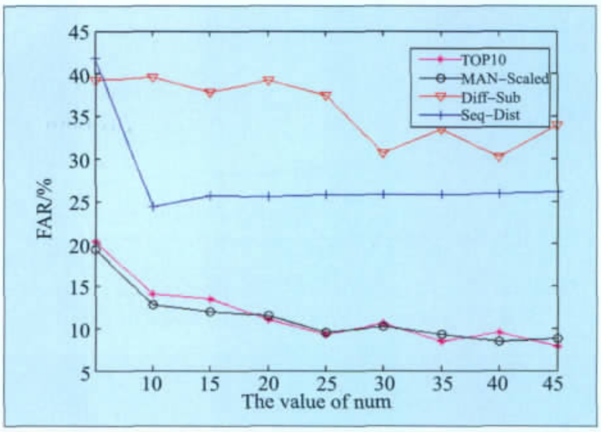

Fig.1 *Variation of FAR with the change of num*

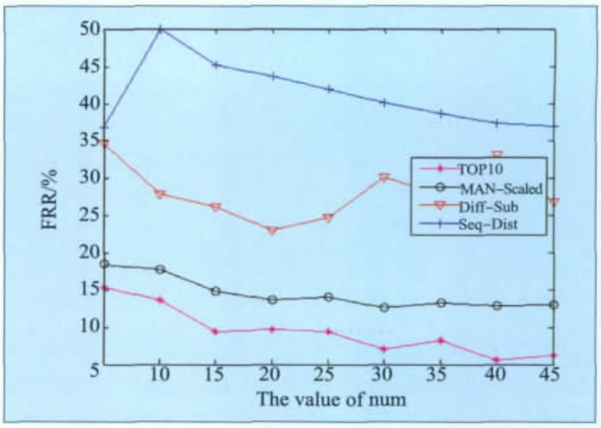

Fig.2 *Variation of FRR with the change of num*

Table I Results of four detectors when num= 45

Detector name	EER	FAR	FRR	FAR + FRR
TOP10	0.072 310 784	0.079 111 765	0.062 254 902	0.141 366 667
Man_scaled	0.108 823 529	0.087 501 961	0.129 901 961	0.217 403 922
Diff_Sub	0.305 392 157	0.339 544 118	0.267 156 863	0.606 700 98
Seq_Dist	not exist	0.260 956 863	0.368 627 451	0.629 584 314

a better performance than the other three. In addition, the recognition accuracy of Man_Scaled and the Diff_Sub does not improve significantly after num=20. Meanwhile, the performance of Seq_Dist has been relatively stable, that is because the detector of Seq_Dist extracts the relative distance as a characteristic value. As the user's familiarity to the keystroke is improved, changes in various keys will also remain relatively stable.

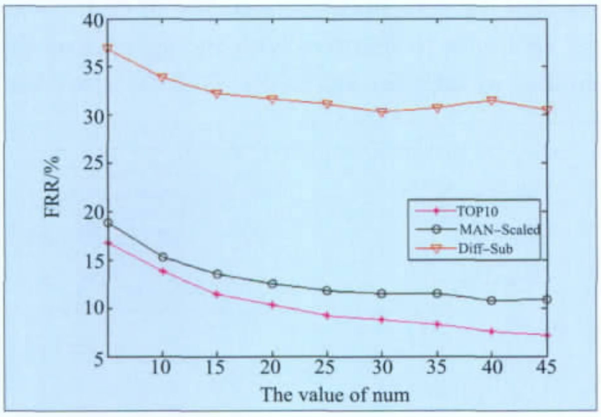

Fig.3 *Variation of ERR with the change of num*

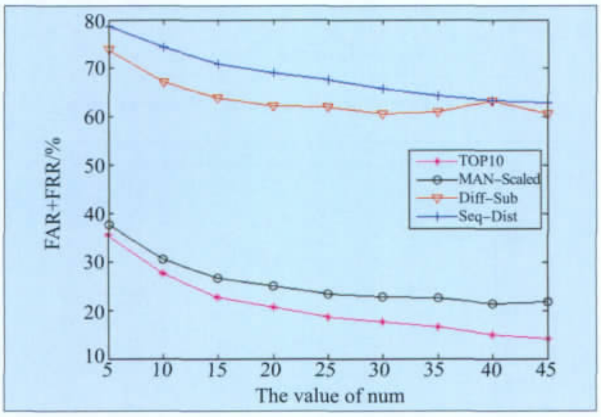

Fig.4 *Variation of FRR +FAR with the change of num*

Finally, we conclude that the smallest value of the sum of FAR and FRR of the detector in this paper is 0.141 4, Man_Scaled is 0.217 4, Seq_Dist is 0.606 7 and Diff_Sub is 0.629 6. So the experimental result shows that the detector in this paper is more efficient than the other three.

V. CONCLUSIONS

This paper presented a methodology that utilized the typing biometrics features to improve the usual login-password authentication. Some experiments were conducted and the better performance was achieved by using a statistical classifier which combined the statistics and the weighted method. The objective in this work is not only to propose a good methodology, but also to utilize a third-party benchmark data set as the testing standard for keystroke dynamics, which means our evaluation methodology could be used by the community to assess new detectors and obtain comparative results. It is because the user's keystroke data plays a crucial role in detection accuracy of various keystroke detectors.

In the further work, in order to improve the accuracy of keystroke authentication recognition, our research work will focus on how to depict the user's feature template in a more efficient way. The future work may also include the update of the feature template so as to improve the user's familiarity of keystroke. 中国通信

Acknowledgements

This paper has been performed in the Project "Key Technology Research of Eavesdropping Detection in the Quantum Security Communication" supported by the National Natural Science Foundation of China under Grant No. 61100205.

References

[1] TEH P, TEOH A, ONG T, et al. Statistical Fusion Approach on Keystroke Dynamics[C]// Proceedings of the IEEE Conference on Signal-Image Technologies and Internet-Based System: December 16-18, 2007, Shanghai. IEEE Press, 2007: 918-923.

[2] LI Jian, JIN Haifei, JING Bo. Improved Security Detection Strategy for Quantum "Ping-Pong" Protocol and Its Security Analysis[J]. China Communications, 2011, 8(3): 170-179.

[3] KANG P, PARK S, HWANG S, et al. Improvement of Keystroke Data Quality Through Artificial Rhythms and Cues [J]. Computers and Security, 2008: 3-11.

[4] AKILA M, SURESH K. Improving Feature Extraction in Keystroke Dynamics Using Optimization Techniques and Neural Network[C]// Proceedings of the International Conference on Sustainable Energy and Intelligent Systems: July 20-22, 2011, Chennai. IEEE Press, 2011: 891-898.

[5] BAZRAFSHAN F, JAVANBAKHT A, MOJALLALI H. Keystroke Identification with a Genetic Fuzzy Classifier[C]// Proceedings of the 2nd International Conference on Computer Engineering and Technology: April 16-18, 2010, Chengdu. IEEE Press, 2010, 4: 136-140.

[6] MARTONO W, ALI H, SALAMI M. Key-Stroke Pressure-Based Typing Biometrics Authentication System Using Support Vector Machines[J]. Computational Science and Its Applications, 2007: 85-93.

[7] SINGH S, ARYA K. Key Classification: A New Approach in Free Text Keystroke Authentication System[C]// Proceedings of the 3rd Pacific-Asia Conference on Circuits, Communications and System, 2011: 1-5.

[8] LI Jian, SONG Danjie, GUO Xiaojing, et al. ID Updating Based RFID Mutual Authentication Protocol for-Low-Cost Tags[J]. China Communications, 2011, 8(7): 122-127.

[9] CHANG T, TSAI C, LIN J. A Graphical-Based Password Keystroke Dynamic Authentication System For Touch Screen Handheld Mobile Devices[J]. Journal of Systems and Software, 2012, 85(5): 1157-1165.

[10] QI Yong,YAO Qingsong, CHEN Ying, et al. Study on RFID Authentication Protocol Theory[J]. China Communications, 2011, 8(1): 65-71.

[11] GUO F. Research on the Security Architecture of LTE/SAE [C]// Proceedings of International Conference ofChina Communication and Information Technology, 2010: 358-362.

[12] WU Yue, YI Ping, LI Jianhua. Security Issues and Solutions in Wireless Communications at Physical Layer[J]. China Communications, 2011, 8(5): 11-19.

[13] MARCIN D, PIOTR P. Overview of Identity Management [J].China Communications, 2008, 5(4): 129-142.

[14] LI Jian, JIN Haifei, JING Bo. Improved Quantum “Ping-Pong” Protocol Based on GHZ State and Classical XOR Operation[J]. SCIENCE CHINA Physics, Mechanics & Astronomy, 2011, 54(9): 1612-1618.

[15] ZHOU Yajian, PAN Anwei, LI Jiguo. An Authenticated Dynamic Key Management Scheme for Clustered Sensor Networks[J]. China Communications, 2010, 7(4): 7-17.

[16] TAN Zouwen. A Provably Secure Identity-based Authentication Multiple Key Agreement Protocol[J]. China Communications, 2011, 8(2): 26-33.

[17] WANG Lifeng, MENG Qinglei, XIAO Chen, et al. Efficient Security Multimedia System on Embedded Platform[J]. China Communications, 2010, 7(4): 120-125.

[18] KILLOURHY K, MAXION R. Comparing Anomaly Detectors for Keystroke Dynamics[C]// Proceedings of the 39th Annual International Conference on Dependable Systems and Networks, 2009: 125-134.

[19] YANG H, LIU J, DAO L, et al. Two-Factor User Authentication Based on Keystroke Dynamics and Password[J]. Information Technology, 2006, 1009-2552(10): 94-97.

[20] LIANG Juan, WANG Xuan, CHEN Weiwei, et al. Recognition of User's Keystroke Features Base on Difference Subspace[J] Computer Engineering, 2007, 33(11): 204-205.

[21] DUDA R, HART P, STORK D. Pattern Classification(2nd edition)[M], John Wiley & Sons, Inc., 2001.

[22] HAIDER S, ABBAS A, ZAIDI A. A Multi-Technique Approach for User Identification Through Keystroke Dynamics [C]// Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, 2000: 1336-1341.

[23] KILLOURHY K, MAXION R(DSN-2009). Keystroke Dynamics-Benchmark Data Set Accompaniment to “Comparing Anomaly-Detection Algorithms for Keystroke Dynamics”. [EB/OL]. http://www.cs.cmu.edu/ ~ keystroke/#sec3.

[24] BLEHA S, SLIVINSKY C, HUSSIEN B. Computer Access Security Systems Using Keystroke Dynamics [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1990, 12(12): 1217-1222.

Biographies

Li Jian, Ph.D., associate professor of Beijing University of Posts and Telecommunications. His research interests include quantum information, quantum computation, quantum communication security, electronic commerce and artificial intelligence. Email: lijian@ bupt.edu.cn

Guo Xiaojing, is a M.S. student in the School of Computer at the Beijing University of Posts and Telecommunications, China. She received B.S. degree in Network Engineer (Information Security) from Zhengzhou University of Light Industry. Her research interests include keystroke Authentication, quantum information, information security, and the security of the Internet of Things. Email: bunny-gxj@163.com

Li Meiyun, is an undergraduate student in the School of Computer at the Beijing University of Posts and telecommunications, China. Her research interests include quantum information, information security, the security of the Internet of things. Email: 1031765038@ qq.com

Li Ruifan, is a lecturer of Beijing University of Posts and Telecommunications, his research interests include quantum information, quantum computation, quantum communication security, information security, artificial intelligence. Email: rfli@bupt.edu.cn