On the Analysis of Keystroke Recognition Performance based on Proprietary Passwords

Alejandro Acien, Javier Hernandez-Ortega, Aythami Morales, Julian Fierrez, Ruben Vera-Rodriguez, Javier Ortega-Garcia

BiDA Lab, Universidad Autonoma de Madrid, Madrid, Spain
{ alejandro.acien, javier.hernadezo, aythami.morales, julian.fierrez, ruben.vera, javier.ortega}@uam.es

Keywords: biometrics, keystroke dynamics, behavior analysis

Abstract

This paper investigates factors affecting the keystroke recognition performance. Keystroke recognition has attracted the interest of scientific community because of the many challenges associated to modelling the typing behavior. Despite the great efforts made during the last decades, the performance of keystroke recognition systems is far from the performance achieved by traditional hard biometrics. This is very pronounced for some users, who generate many recognition errors even with the most sophisticate recognition algorithms. Our purpose here is to study factors affecting the performance of users for scenarios in which each user employ a proprietary password based on familiar information. The experiments comprise a public database with 300 users (300 passwords) and four state-of-the-art recognition systems recently evaluated during the Keystroke Biometrics Ongoing Competition. The results suggest the importance of the correct alignment of samples and intra-class variability despite the impact of the length of the password and timing features.

1 Introduction

Behavioral biometrics became popular during last decades because of their ease of use and large number of applications including user authentication (e.g., ID management), user profiling (e.g., gender or age prediction), health (e.g., neuromotor diseases detection), among others. The behavioral biometrics analyze "something that we do" instead of classical physiological biometrics which analyze "something that we are" [1]. Some of the most popular behavioral biometrics are speech, handwriting, gait and keystroke.

Biometric recognition systems validate the subject identity by comparing the subject template (pre-stored in a database) with a captured biometric sample. Keystroke biometrics refers to technologies developed for automatic user authentication/identification based on the classification of their typing patterns [2]. These technologies present several challenges associated to modeling and matching dynamic sequences with high intra-class variability (e.g., samples from the same user show large differences), low inter-class variability (e.g., samples from different subjects show similarities) and variable performance (e.g., human behavior is strongly user-dependent and varies significantly between subjects).

From the industry's point of view, keystroke technologies offer authentication systems capable of improving the security and trustworthiness of web services (e.g., banking, mail), digital contents (e.g., databases) or new devices (e.g., smartphones, tablets). Given the wide range of potential practical applications mentioned above, a heterogeneous community of researchers from different fields has produced in the last decade a very large number of works studying different aspects of keystroke recognition. Those contributions have been compiled in several surveys [2,3,4,5] that describe the technology in terms of performance, databases, privacy and security. The techniques are usually divided into **fixed text** (the text used to model the typing behavior of the user and the text used to authenticate is the same) and **free text** (the text used to model the typing behavior and the text used to authenticate do not necessarily match). For the rest of this work we will focus on fixed text scenario

The performance of keystroke biometrics systems is strongly dependent on the application (e.g., fixed or free text) and databases (e.g., different users show very different performances). Moreover, the performance of keystroke users is difficult to predict [6]. There is a large margin between performance of different users and it is possible to observe users with performances ten time worse than others independently of the keystroke authentication systems employed. The reasons of this variable performance have attracted the interest of researchers [6,7,8,9]. In [6] researchers explored different ways to predict the performance of good and bad users (i.e., users with lowest and highest error rates respectively). The authors found that it is possible to ascertain the performance of users using exclusively the genuine samples and the Kullback-Leibler divergence between their features. In [7,8] researchers explored the stability of the patterns associated to the keystroke rhythms. They found that users need several repetitions to stabilize their typing behavior and a strong relationship between performances and length of the password (longest passwords produce lowest error rates). The impact of the complexity was evaluated in [9] in which authors proposed a complexity index calculated for each password. Their results suggest that complex passwords improve the performance of keystroke recognition systems. A common drawback of these studies is the use of unique-

Table 1. Left: Performance (baseline) per user for all systems. Right: EER averaged for good and bad users. The threshold calculated to discriminate between both groups was 10% EER for all systems.

System	Mean EER (%)
P1	11.26
P2	8.81
P3	14.36
P4	4.62

System	Mean EER (%)	
	Good Users	Bad Users
P1	6.13	25.35
P2	4.57	24.78
P3	5.54	24.82
P4	3.07	23.91

password assumption (e.g., "*.tie5Roanl*" [6,7,10] "*try4-mbs*" [7,11] and "*greyc laboratory*" [7,12]) for all subjects in the database. In real applications, the most likely scenario is the one in which each user has a proprietary password different to the other users. A secondary limitation of these studies is that databases rarely surpass one hundred users.

In this work we extend the previous studies by: i) analyzing different factors that affect the keystroke recognition performance for scenario in which each user type a **proprietary password** (300 passwords); ii) we employ one of the largest databases available with **300 users** acquired in 4 different sessions and **four state-of-the art algorithms** recently evaluated during the Keystroke Biometrics Ongoing Competition; iii) we provide **new insights on keystroke recognition** performances including results that contradict what has been known to date about the length of the passwords and its performances.

The rest of the paper is organized as follows. Section 2 describes the database used in this work. Section 3 reports the experiments and results while Section 4 summarizes the conclusions.

2 Keystroke Ongoing Competition Database

The database used in this work is the Keystroke Biometrics Ongoing Competition (KBOC) database [13,14]. The dataset is composed of keystroke sequences from 300 subjects acquired in four different sessions distributed in a four months time span. Thus, three different levels of temporal variability are taken into account: (i) within the same session (the samples are not acquired consecutively), (ii) within weeks (between two consecutive sessions), and (iii) within months (between non-consecutive sessions). Each session comprises 4 case-insensitive repetitions of the subject's name and surname (2 in the middle of the session and two at the end) typed in a natural and continuous manner. Note that passwords based on name and surname are very familiar sequences that are typed almost on a daily basis. This allows us to reduce the intra-class variability and to increase the inter-class variability.

The database was captured in a university environment, being the vast majority of acquired subjects proficient in the use of computers and keyboards. No mistakes are permitted (i.e., pressing the backspace), if the subject gets it wrong, he/she is asked to start the sequence again. The names of three other subjects in the database are also captured as forgeries, again with no mistakes permitted when typing the sequence. However, during the acquisition we observed that around 4% of samples (equally distributed among genuine and impostors) present inconsistencies that produce different lengths in the sequences. The use of shift keys can vary the number of keys pressed even if the final result does not change. For example, the sequences Shift+Shift+a=A and the sequences Shift+a=A have different lengths but same text as output. We consider these samples as matching and therefore they are part of the database employed for the competition. The time (in milliseconds) elapsed between key events (press and release) is provided as the keystroke dynamics sequence. Imitations are carried out in a cyclical way, i.e., all the subjects imitate the previous subjects, and the first one imitates the last subjects.

The experimental protocol used in this work is the same proposed during the KBOC Competition. It is based on the following steps, for each user: (i) Participants have 4 training samples (genuine samples from the 1st session) as enrolment data. (ii) 20 test samples (genuine and impostor samples randomly selected from remaining samples not used for training) are used to evaluate the performance of the systems. The number of genuine and impostor samples per user varies between 8 and 12 (but the sum is equal to 20 for all of them). This variable number of genuine and impostor samples helps to avoid algorithms that exploit cohort information. (iii) Each test sample is labelled with its corresponding user model and performance is evaluated according to the verification task (1:1 comparisons). The performance is evaluated in form of User-dependent Equal Error Rate EER. EER refers to the value where False Match Rate (FMR, percentage of impostors users classified as genuine) and False Non-Match Rate (FNMR, percentage of genuine users classified as impostors) are equal. The EER has been calculated independently for each of the 300 subjects (300 different decision thresholds). EER is the average individual EER from all subjects.

3 Experiments and Results

3.1 Methodology

We will analyze the performance of 4 state-of-the-art keystroke recognition systems evaluated during the KBOC Competition [11,14]. The systems were submitted by 4 different participants. We have chosen the best system from each participant among the 31 systems submitted during the competition (see [14] for details). Table 1 (Left) shows the performance of all 4 systems according to the experimental protocol proposed. This performance will be used as baseline

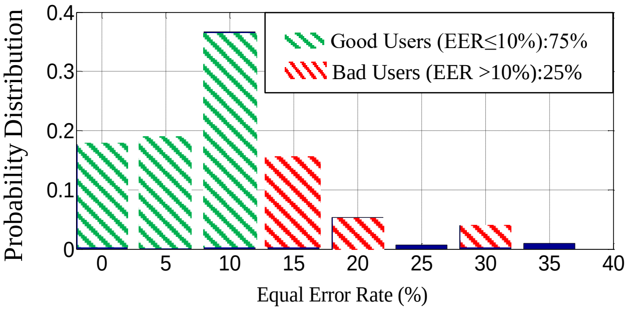

Fig 1. Probability distribution of Equal Error Rate (averaged from all 4 systems) among the database population.

for the rest of the experiments. The results show a large difference between the performance of the Participant 4 (P4) and the rest of participants. The largest differences between participants lie in the pre-processing (sequence alignment and feature normalization) and post-processing techniques (score normalization) applied. The score normalization applied by P4 allows reducing the EER up to 4.62%. The next sections will analyze different factors affecting the performance of keystroke recognition systems at three levels: Classification level (by analyzing the scores obtained by the systems), Feature level (by analyzing the features used as input for the systems) and Score level (by analyzing techniques used for score normalization).

3.2 Results: Performance analysis at classification level

Baseline: The performance of keystroke dynamics is strongly user-dependent. As an example, Fig. 1 shows the probability distribution of the EER (averaging the performance of all 4 systems) obtained independently for each of the 300 users. The results show a large margin between performances of different users (from 0% to 35% of EER). In addition, it is remarkable the large number of users with 0% of EER for all 4 systems (around 20% of users). What are the main factors affecting this large difference between performances obtained for each user is the final aim of this work.

Good vs. Bad Users: In order to analyze the performance of users, the database was divided into two groups (independently for each system) attending to the EERs of the users. Users with lower EER (EER≤10%) were named as good users while users with higher EER (EER >10%) where named as bad users, the average of the EER for each group are summarized in Table 1 (Right). While good users show mean EER ranging between 3% and 6%, the bad users show up to 25% mean EER. The good users represent around 75% of the database while bad users the remaining 25%. The probability distribution of classification scores from test samples (normalized between 0-1 for all 4 systems) can be seen in Fig. 2 (Left). The distributions shown that overlap between both genuine and impostor scores is greater for bad users as is expected. However, the degradation of the genuine scores is higher, suggesting that intra-class variability (difference between samples of the same user along different sessions) is more important than the inter-class variability (ability of the impostor) in this scenario. Table 2 shows confusions matrices for both groups and all 4 systems. The average percentage of coincidence between good users is 55% and 70% for bad ones. The superior percentage of bad users suggests that worst users are difficult to identify for all 4 systems. On the other hand, there are 30% and 45% of bad and good users respectively that were classified into different quality groups depending of the system. These results suggest a large complementarity between systems (i.e., users with bad performances for system A can show good performances for system B).

3.3 Results: Performance analysis at feature level

Length of the Password: The first experiment is based on the idea that the length of the passwords can affect the performance of the systems [7,9]. Long passwords can be better for discriminate between impostors and genuine users due to it carries more biometric user information. However, the results showed in Fig. 2 (Right) suggests there is not dependence between length of the passwords and system performance. These results contradict previous works [7,9] which states clear differences between performances obtained by long and short passwords. There are two main reasons to explain these results: i) passwords used in this KBOC database are composed by familiar words (name and surname) instead of alphanumeric sequences of symbols (e.g., “tie5Roanl” and “try4-mbs”). The users of KBOC database show very stable features as they type very familiar sequences; ii) the length of the passwords in KBOC database ranges between 12 and 38 symbols while previous studies were based in passwords with a maximum length of 16 symbols. Based on our experiments and scenarios, he length of the password is not a key factor which determine the keystroke performance.

Timing: Regarding two of the most popular characteristics on keystroke dynamics, we calculated the values of Hold Time (difference between timestamps of press and release events of

Table 2. Confusion matrix for good users (Left) and bad users (right). System P4 (row 4) has the largest number of good users in comparison with the rest others systems.

	P1	P2	P3	P4
P1	100	48.46	42.47	41.97
P2	72.10	100	57.73	54.92
P3	36.05	32.56	100	30.05
P4	94.18	82.17	79.50	100

	P1	P2	P3	P4
P1	100	85.96	75.77	95.33
P2	68.69	100	61.67	78.50
P3	80.37	81.86	100	85.98
P4	47.66	49.12	40.53	100

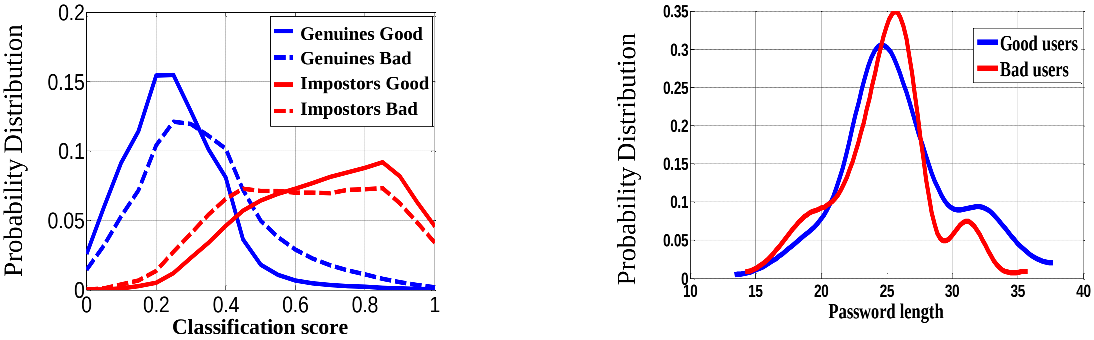

Fig 2. Probability distributions of classifications scores (Left) and length of passwords (Right) for good and bad users (curves averaged from all four systems).

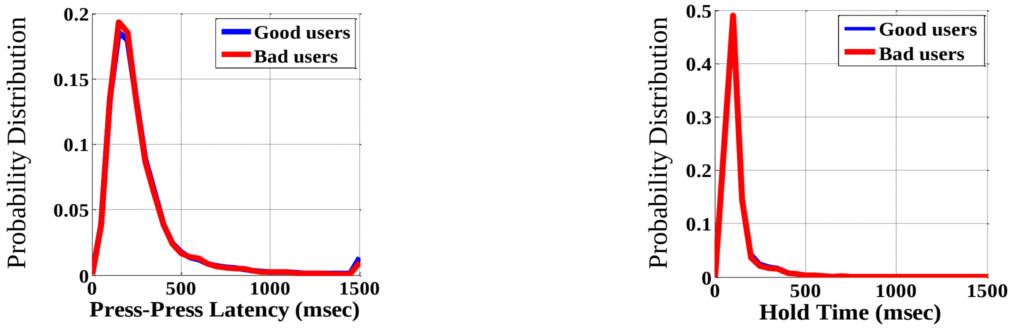

Fig 3. Probability distribution of features for good and bad users (curves averaged from all four systems).

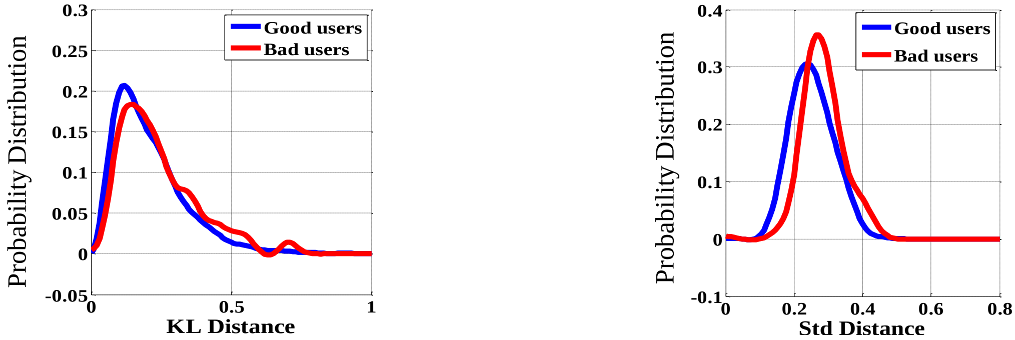

Fig 4. Probability distribution of enrolment set variability (measured in form ok Kullback-Leibler divergence and standard deviation) for good and bad users (curves averaged from all four systems).

the same key) and Press-Latency (difference between timestamps of press and press events of consecutive keys) for each user. Fig. 3 shows both features for good and bad users and any difference between them have been appreciated. Good and bad users show exactly the same distributions of time. This result suggests that there are no differences in terms of time features (i.e., time between individual key events).

Misalignment: Around 4% of the samples in the database have different number of keys pressed (mainly because of the use of the shift keys). These sequences may produce misalignments during the comparison of training and test samples and performance degradation up to 300% (see [13] for details). The number of misaligned samples in bad users is two times greater than good users. These results suggest that the correct alignment of sequences is critical for keystroke recognition performance.

Stability of the Features: For this experiment we measured the distance between training samples and genuine test samples for each user. In order to measure the distance, we propose two methods: standard deviations (std) and Kullback-Leibler divergence (KL). For each test sample, both distances are calculated as the distance between the test features and the enrolment feature vector (calculated averaging the 4 training feature vectors). The Fig. 4 shows distances for good and bad users, KL distance seems to be very similar for both groups but small differences in std distance were observed. This difference in std distance suggest that good users tend to have less keystrokes variations.

Table 3. EER for all systems with (EER'$_G$) and without (EER$_G$) score normalization. In brackets we show the improvement.

System	EER$_G$	EER'$_G$
P1	15.7	12.05 (↓23%)
P2	11.83	9.08 (↓23%)
P3	17.95	14.54 (↓19%)
P4	20.15	5.31 (↓73%)

3.4 Results: Performance analysis at score level

EERs showed in Table 1 (Left) were calculated independently for each of the 300 subjects (300 different decision thresholds), these EERs are calculated as the average of the individual EER from all subjects [3,10,12]. To analyse the impact of the score normalization in the performance, average EER from the whole database (using only one decision threshold for all users) are summarized in Table 3. To differentiate between booth types of EER, EER$_G$ denote average from whole database while EER denote average at user level. Three different techniques of score normalization are proposed for this experiment with similar results: min-max, mu-sigma and tangh (see [15] for details). The best performance was achieved with a relative min-max normalization technique proposed in [16] and described below:

$$score' = \frac{score - min_i}{max_i - min_i} \tag{1}$$

where:

$$min_i = \mu_i - 2 \times \sigma_i \tag{2}$$

$$max_i = \mu_i + 2 \times \sigma_i \tag{3}$$

these μ_i and σ_i are the mean and standard deviation of the user i obtained from the 20 test scores available for each user (optimist a posteriori normalization).

Table 3 shows a significant improvement for all system when score normalization is applied. The experiment show that score normalization can be used to improve performance by 20%. System P4 had the largest improvement ranging from 20.15% to 5.31% EER. These results suggest that the strong impact of the normalization techniques in the performances. Note that best results are obtained using normalization parameters (mean and std of EER'$_G$) optimized according to the scores of each user. In some applications the scores available to model each user are limited and other strategies should be explored.

4 Conclusions

This paper has analyzed different factors affecting the performance of biometric recognition based on keystroke dynamics. The performance of keystroke dynamics is highly user-dependent and it is usual to find large performance deviations among users even with the most competitive recognition algorithms. We have analyzed the performance of four state-of-the-art keystroke recognition systems (summarized in Table 4). The experiments suggest that: i) the length of the password does not affect the performance of keystroke authentication for long passwords (>12 symbols) and familiar sequences; ii) intra-class variability has higher influence than inter-class variability; iii) misaligned samples have a strong impact on the performance; iv) the timing features from good and bad users are similar, v) score normalization techniques offers a huge improvement for algorithms with good intra-class adaptation but does not represent a realistic scenarios where a few training samples are available for these techniques.

Acknowledgements

This work was funded by the project CogniMetrics (TEC2015-70627-R) and JdC contract (JCI-2012-12357) from MINECO/FEDER.

References

[1] S. Prabhakar, S. Pankanti, A. K. Jain. "Biometric Recognition: Security and Privacy Concerns", *IEEE Security & Privacy*, pp. 33-42, (2003).

[2] A. Peacock, X. Ke, M. Wilkerson. "Typing patterns: A key to user identification", *IEEE Security and Privacy*, **vol. 2, no. 5**, pp. 40-47, (2004).

[3] Y. Zhong, Y. Deng. A survey on keystroke dynamics biometrics: approaches, advances, and evaluations. In: Y. Zhong, Y. Deng (eds.). "Recent Advances in User Authentication Using Keystroke Dynamics Biometrics", *Science Gate Publishing*, pp. 1-22, (2015).

Table 4. Summary of the impact (↑ low, ↑↑ medium and ↑↑↑ high) for each factor based on our experimentation in keystroke dynamics for KBOC database.

Factors	Performance	Usability	Computational cost
Length	↑	↑↑↑	↑
Timing	↑	↑	↑↑
Misalignment	↑↑↑	↑↑	↑↑↑
Stability	↑↑	↑	↑↑
Normalization	↑↑↑	↑	↑↑↑

[4] M. L. Ali, J. V Monaco, C. C Tappert and M. Qiu, “Keystroke Biometric Systems for User Authentication”, *Journal of Signal Processing Systems*, pp. 1-16, (2016).

[5] D. Shanmugapriya and G. Padmavathi, “A survey of biometric keystroke dynamics: approaches, security and challenges”, *Int. Journal of Computer Science and Information Security*, **vol. 5, no. 1**, pp. 115–119, 2009.

[6] A. Morales, J. Fierrez, J. Ortega-Garcia. “Towards predicting good users for biometric recognition based on keystroke dynamics”. *Proc. of European Conf. on Computer Vision Workshops*, Springer LNCS-8926, pp. 711-724, (2014).

[7] J. Montalvão, E. O. Freire, M. A. Bezerra Jr. and R. Garcia. “Contributions to empirical analysis of keystroke dynamics in passwords”, *Pattern Recognition Letters*, **vol. 52, no. 15**, pp. 80-86, 2015.

[8] Z. Syed, S. Banerjee, Q. Cheng, B. Cukic. “Effects of user habituation in keystroke dynamics on password security policy”, *IEEE Computer Society*, pp.352–359, (2011).

[9] S. Mondal, P. Bours, S. Z. S. Idrus. “Complexity Measurement of a Password for Keystroke Dynamics: Preliminary Study”. *Proc. Of The 6th Int. Conf. on Security of Information and Networks*, pp. 301-305, (2013).

[10] K. S. Killourhy and R. A. Maxion. "Comparing Anomaly Detectors for Keystroke Dynamics", *Proc. of the 39th Ann. Int. Conf. on Dependable Systems and Networks*, Estoril, Lisbon, Portugal, IEEE CS Press, pp. 125-134, (2009).

[11] C. C. Loy, C. P. Lim, W. K. Lai. “Pressure-based typing biometrics user authentication using the fuzzy ARTMAP neural network”. *Proc. of the 12th Int. Conf. on Neural Information Processing*, pp.1–6, (2005).

[12] R. Giot, M. El-bed and R. Christophe. “Greyc keystroke: a benchmark for keystroke dynamics biometric systems”, *Proc. of IEEE Int. Conf. on Biometrics: Theory, Applications and Systems*, Washington DC, pp. 1-6, (2009).

[13] A. Morales, J. Fierrez, R. Tolosana, J. Ortega-Garcia, J. Galbally, M. Gomez-Barrero, A. Anjos, S. Marcel. “Keystroke Biometrics Ongoing Competition”, *IEEE Access*, 4, pp. 7746-7746, (2016).

[14] Morales, *et al*. “KBOC: Keystroke Biometrics OnGoing Competition”. *Proc of The IEEE Eighth Int. Conf. on Biometrics: Theory, Applications, and Systems*, pp. 1-6, (2016).

[15] Robert. “Large-Scale Evaluation of Multimodal Biometric Authentication Using State-of-the-Art Systems”, *Proc of The IEEE Transactions on Pattern Analysis and Machine Intelligence*, **vol. 27, no. 3**, (2005).

[16] J. V Monaco. “Robust Keystroke Biometric Anomaly Detection”, *arXiv preprint*, pp. 1-7, 2016.