KBOC: Keystroke Biometrics OnGoing Competition

Aythami Morales[1], Julian Fierrez[1], Marta Gomez-Barrero[1], Javier Ortega-Garcia[1], Roberto Daza[1], John V. Monaco[2], Jugurta Montalvão[3], Jânio Canuto[3], Anjith George[4]

[1]ATVS-Universidad Autonoma de Madrid, C\ Francisco Tomás y Valiente, 11, 28049, Madrid, Spain
[2]U.S. Army Research Laboratory, Aberdeen Proving Ground, USA
[3]UFS - Universidade Federal de Sergipe, 49100-000, São Cristóvão, Brazil
[4]Indian Institute of Technology Kharagpur, Kharagpur 721302 , West Bengal, India

{aythami.morales, julian.fierrez, marta.barrero, javier.ortega}@uam.es, john.v.monaco2.ctr@mail.mil, jmontalvao@ufs.br, anjith2006@gmail.com

Abstract

This paper presents the first Keystroke Biometrics Ongoing evaluation platform and a Competition (KBOC) organized to promote reproducible research and establish a baseline in person authentication using keystroke biometrics. The ongoing evaluation tool has been developed using the BEAT platform and includes keystroke sequences (fixed-text) from 300 users acquired in 4 different sessions. In addition, the results of a parallel offline competition based on the same data and evaluation protocol are presented. The results reported have achieved EERs as low as 5.32%, which represent a challenging baseline for keystroke recognition technologies to be evaluated on the new publicly available KBOC benchmark.

1. Introduction

Biometric recognition is a wide research area which includes researchers from pattern recognition and machine learning communities. Biometric technologies are usually divided into physiological (*e.g.* fingerprint, face, iris) and behavioral (*e.g.* signature, gait, keystroke) according to the nature of the biometric trait used. Behavioral biometrics have attracted the interest of researchers and industry because of its ease of use, transparency and large number of potential applications [1].

Keystroke biometric applications have been investigated over the past several decades, attracting both academics and practitioners. These technologies present several challenges associated to modeling and matching dynamic sequences with high intra-class variability (*e.g.* human behavior is strongly user-dependent and varies significantly between subjects). In addition, the simple nature of the data (time sequences) makes keystroke biometrics a good field to introduce new researchers (without previous experience on biometric applications) in this challenging area.

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). The keystroke recognition community is heterogeneous and includes researchers from different disciplines [1][2]. The number of algorithms and approaches is large and it is difficult to establish a baseline. As a behavioral biometric trait, 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). Public benchmarks have been proposed, offering the opportunity to compare systems under the same conditions. Some of the most popular keystroke benchmarks based on fixed-text sequences are CMU [3], GREYC [4], MIMOS [5], Clarkson [6], BeiHang [7] and the recently published ATVS-Keystroke [8]. Even though these benchmarks represent valuable resources, they suffer from two important limitations: the small number of subjects (no more than 133 subjects) and their application scenario, which assumes that all users share the same password in most of the cases. In real applications, the most probable scenario is the one in which different users have different passwords.

To the best of our knowledge, there is only one previous keystroke recognition competition that was hosted during the IAPR International Conference on Biometrics 2015: "*One-handed Keystroke Biometric Identification Competition*" [9]. In this competition, keystroke technologies were evaluated in a free-text scenario involving the response of 63 students to three online exams. The competition analyzed the performance of person authentication algorithms under challenging conditions, in which users were forced to type using only one hand instead of a more natural way, using two hands.

Traditional biometric competitions give a static snapshot of the state-of-the-art in a specific research area. The main

problem is how to encourage researchers to invest their resources and time to participate in these competitions (usually operative during a short window of time). Without the participation of the main players, the snapshot will be inaccurate. In contrast, the ongoing competitions provide a dynamic view constantly updated by the community. The FVC-onGoing competition [10] is a successful example with more than 900 participants and more than 4000 algorithms evaluated since 2009 for fingerprint technologies.

The keystroke competition described in the present work tries to complement the previous experiences by: (i) proposing the first keystroke **ongoing competition** which overcomes the limitations of traditional competitions based on a static snapshot of the state-of-the-art; (ii) disclosing a public benchmark involving 7600 keystroke sequences from **300 users**, simulating a realistic scenario in which each user types his own sequence (given name and family name) and impostor attacks (users who try to spoof the identity of others) and (iii) being an online competition carried out over a **fully reproducible** framework based on the BEAT platform[1] [11] and an offline competition as baseline.

The rest of the paper is organized as follows: Section 2 describes the database and evaluation protocols. Section 3 presents the best systems submitted by the participants to the offline competition. Section 4 reports the experiments and results. Finally, Section 5 summarizes the conclusions.

2. Dataset and Protocols

2.1. Dataset and Evaluation Protocol

The dataset proposed for the competition is part of the BiosecurID multimodal database [12] and consists of keystroke sequences from 300 subjects acquired in four different sessions distributed in a four month 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. 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, the use of shift key produces sequences (around 10% of samples equally distributed among genuine and impostors) with different number of keys pressed even for the same text typed. For example the sequences *Shift*+*Shift*+*a*=*A* and the sequences *Shift*+*a*=*A* have different lengths but same text as output. 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 imitate the last subjects. The main statistics of the dataset proposed for the competition are summarized in Table 1.

Table 1. Summary of the main statistics of the database proposed for the competition.

Characteristics	#
Number of users (Testing Set)	300
Number of users (Development Set)	10
Number of sessions	4
Training samples per user	4
Test samples per user	20
Genuine samples per user*	8-12
Impostor samples per user*	8-12
Total genuine comparisons	3028
Total impostor comparisons	2972
Average separation between sessions	1 month
Average length of the key sequence	25.55

*In order to increase the difficulty, the number of genuine and impostor samples per user varies depending on the user. Participants do not know this number.

The test samples remained sequestered (*i.e.*, participants did not know whether they are genuine or impostors samples). In addition, a small development set (10 users with labeled samples) and baseline algorithms were provided to the participants.

The experimental protocol was based on the following steps, for each user: i) participants have 4 training samples (genuine samples from the 1st session) as enrollment data; ii) 20 test samples (genuine and impostor samples randomly selected from the 24 samples available from 2nd to 4th sessions) 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); iii) each test sample is labeled with its corresponding user model and performance is evaluated according to the verification task (1:1 comparisons).

There are two modes of participation: ongoing and offline. Dataset and evaluation protocols of both modes of participation are exactly the same. The performance of the offline evaluation (detailed in Section 4) will be used as baseline for the ongoing competition.

2.2. Ongoing Competition

The competition exploits the potential of the BEAT platform, which was created under the FP7 EU BEAT project to promote reproducible research in biometrics. The

[1] https://www.beat-eu.org/platform/

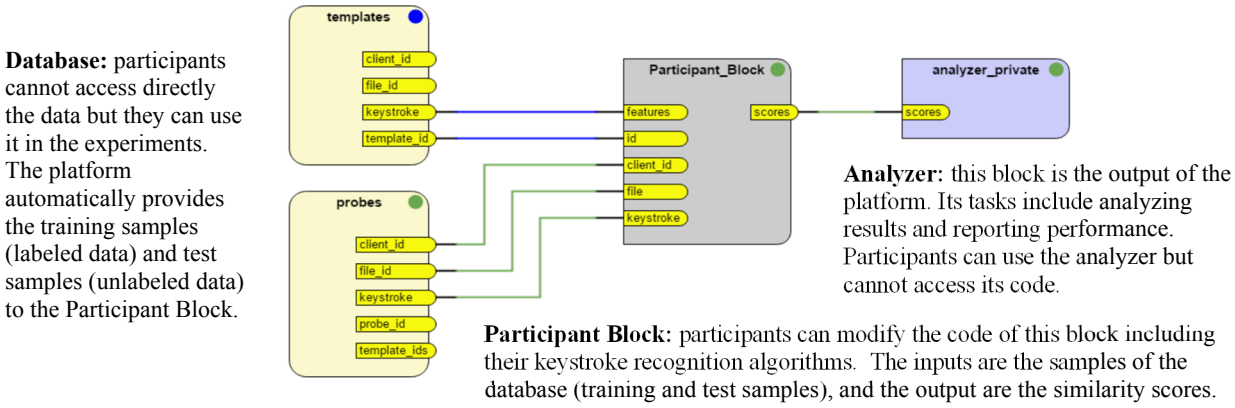

Figure 1. Toolchain of KBOC developed on BEAT (https://goo.gl/8DJQN7)

BEAT platform is a European computing e-infrastructure for Open Science proposing a solution for open access, scientific information sharing and re-use including data and source code while protecting privacy and confidentiality. The platform is a web-application allowing experimentation and testing in pattern recognition.

KBOC provides the data and modules necessary to run the evaluation and the BEAT platform ensures that the system is correctly executed, also providing the results. Different algorithms and systems can be easily compared. Figure 1 shows the toolchain of KBOC with the modules involved in the evaluation. The platform also provides an attestation mechanism for the reports (*e.g.*, scientific papers, technical documents or certifications). The competition website provides instructions and examples to facilitate the participation of researchers without previous experience on BEAT. There is no limit regarding the number of systems evaluated, and the results are automatically provided to the participants on the platform (*i.e.*, the performance of the systems is available in real time). The platform will be available beyond the offline competition and it is a new valuable resource for the keystroke recognition community (see the website [2] for details, tutorial and extra material). The ongoing platform is available online [3] as well as results of all systems evaluated[4].

It should be noted that participating in the ongoing competition and using of the platform does not imply the publication of the code and confidentiality is in any case granted. The organizers have no access to the private code evaluated by the platform but only to the results obtained. Reproducibility is granted by allowing execution permission without code access, thereby preserving confidentiality.

2.3. Offline Competition

In addition to the ongoing evaluation platform, a traditional offline competition was proposed to promote reproducible research serving as baseline for the ongoing evaluation. The training set and test set (described in *Section 2.1*) were available at KBOC website. The keystroke recognition algorithms were executed at the participant premises according to the competition protocol. The scores (comparisons between user models and genuine/impostors samples) obtained by the participants were sent to the KBOC organization. To avoid overfitting, the number of submissions was limited to 15 different systems that were evaluated after the submission deadline.

3. Description of Participating Systems

There was a total of 12 institutions from 7 different countries registered for the competition (5 from USA, 2 from India and 1 from Norway, Argelia, The Netherlands, Brazil and China). Four from the registered institutions finally submitted their systems for a total number of different systems evaluated equal to 37. This section presents the descriptions of the three best systems evaluated during the offline competition.

3.1. U.S. Army Research Laboratory

System 6 submitted by ARL used a Manhattan distance anomaly detector with keystroke duration and press-press (PP) latency features. While Manhattan distance generally yields relatively low EER among distance-based anomaly detectors [3], the low EER can also be attributed to preprocessing, feature normalization, and score normalization.

The raw data of each sample was first converted to a sequence of keystroke events with each event described by the key, press time, and duration. Although the data

[2] https://sites.google.com/site/btas16kboc/
[3] https://goo.gl/VsKgVM
[4] https://goo.gl/i7M5n5

collection procedure of the test set did not allow mistakes or backspace by the genuine users, a variety of keystrokes in both template and unknown samples could be observed for each claimed identity. A simple algorithm was developed to establish a correspondence of features between samples for each claimed identity.

The target keystroke sequence was selected as the minimum length sequence in the template samples. In the case where there were multiple minimum length sequences that differed by a permutation, the target sequence was selected randomly. A modified dynamic time warping (DTW) algorithm then matched the keystrokes of every other sample to the length-M target sequence sorted by press time. The M key-hold durations and $M-1$ PP latencies were then extracted from both the template and query samples according to the keystrokes in the target sequence. While the PP latency features for the target sequence were strictly positive, the PP latency features for the other samples were negative for permuted keystrokes. The duration and PP latency features of each claimed identity were then normalized to within one standard deviation (SD) of the mean duration and mean PP latency, respectively, of the genuine samples.

Following feature extraction, the Manhattan distance to the mean template feature vector was calculated. The distances from unknown samples to each claimed identity were then normalized to within ±2 SD of the mean, with distances outside that range clipped to $[0,1]$. This procedure yielded an EER of 6.95 ± 1.17% on the development set, obtained through a Monte Carlo validation procedure. Following the reproducibility criteria of the competition, the code is available at[5].

3.2. Universidade Federal de Sergipe

The UFS team self-imposed three restrictions in order to properly simulate an actual biometric system:

- R1: The number of test samples per subject is not known beforehand.
- R2: The proportion of genuine and impostor amongst the test samples are unknown.
- R3: Sequential test samples simulate system interrogation through time, therefore a score obtained at a given time cannot be used to improve previous scores.

As in [13], Press-Press (PP) and Hold-time (H) time intervals were equalized with parameters $\mu_{\mathrm{PP}} = -1.61$, $\sigma_{\mathrm{PP}} = 0.64$, $\mu_{\mathrm{H}} = -2.46$ and $\sigma_{\mathrm{H}} = 0.33$ respectively, through a non-linear mapping:

$$y = \frac{1}{1 + \exp\left(-\frac{1.7(\log_{\mathrm{e}}(x) - \mu)}{\sigma}\right)} \qquad (1)$$

where x stands for a time interval (in seconds).

For each enrolled subject, a set of PP and H vectors were taken as user templates. During interrogation, the unlabeled PP vector was compared to a single minimum profile obtained through the so-called shuffling procedure (Bleha, 1988), for it is slightly better than using mean profile, yielding d_{PP}. Likewise, d_{H} stands for the minimum distance between templates and the tested H vector. All distances were sums of absolute differences (Manhattan distance) divided by the length of the sample. In case of inconsistent vector lengths, the shorter one was compared to each sub segment of the longer one and the minimum distance is kept (*i.e.*, only time intervals are considered, not character mismatches). Final test score was computed as:

$$d = 0.75 d_{\mathrm{PP}} + 0.25 d_{\mathrm{H}} \qquad (2)$$

Moreover, the template set was automatically appended with new samples every time a score lower than 0.14 is found, thus influencing future scores (*i.e.* online template adaptation).

We highlight that it is possible to improve performance if no restrictions are imposed, but enrollment-interrogation simulation would be less realistic. For instance, by using the training set, through 50 independent runs (*i.e.* independent choices of 4 training signatures per run), an average EER of 8.0% ± 1% (standard deviation) was obtained. By contrast, if the restriction R3 is violated and an *a posteriori* score normalization is done, the EER drops to 6.5 % ±1%, for the same system.

3.3. Indian Institute of Technology Kharagpur

In this approach, the time intervals between consecutive key events are used as the feature vectors. The raw data obtained consisted of a sequence of N keys. Let M be the number of key events (key-press and key-release events) in the sequence. The feature vector was modelled based on the time between two consecutive key events irrespective of press or release events. The formulation is described below.

Let the feature vector of a test sample be, $\mathbf{f} = [f_1, f_2, \dots, f_M]$, where, f_i is the time interval between $(i-1)^{th}$ and i^{th} key events, where $i = 1,2,3, \dots, M$.

Similarly consider the enrolment set $\{\mathbf{g}^k\}_{k=1}^{T}$, where $\mathbf{g}^k = [g_1^k, g_2^k, \dots, g_M^k]$ $k \in 1, \dots, T$, with $T = 4$ samples and M the number of features for each keystroke sequence. Two distance measures were computed between the feature vector $\mathbf{f}$ of the test sample and the enrollment set $\{\mathbf{g}^k\}_{k=1}^{T}$. The two distance metrics used find the absolute distance to the nearest neighbor in each feature dimension independently. A combination of mean and median of these distances was used as the final distance metric. The distance measures were computed as:

$$\mathbf{AD}(i,k) = \left|g_i^k - f_i\right|, \quad k = 1, \dots T \quad \text{and} \quad i = 1,2,3, \dots, M \qquad (3)$$

[5] https://github.com/vmonaco/kboc

$$\mathbf{md}(i) = \min_{k \in [1,\dots T]} ad_{ik}, \qquad i = 1,2,3,\dots,M \qquad (4)$$

where ad_{ik} is an element of matrix $\mathbf{AD}(i,k)$ and the final distance was obtained as:

$$d = \text{mean}(\mathbf{md}) + \text{median}(\mathbf{md}) \qquad (5)$$

4. Results

This section presents the final results of the offline competition while the ongoing results can be seen at the BEAT platform[6]. As it is an ongoing competition, the results will be automatically updated with any new submission.

Regarding the offline competition, the participants were allowed to submit up to 15 different systems before the deadline. As previously mentioned, the algorithms were compared after the deadline, thus being the performance of all systems reported after the submission period ended, according to the following indicators:

- Global Equal Error Rate (EER_G): unique EER calculated using all genuine and impostor scores and only one threshold for all users.
- User-dependent Equal Error Rate (EER_U): the EER is calculated independently for each of the 300 subjects (300 different thresholds). EER_U is the average individual EER from all subjects. This EER is common in the keystroke dynamics literature [2][3][4].
- FMR_{100}: the lowest False Non-Match Rate for False Match Rate equal to 1%.
- Detection-Error Tradeoff (DET) curve: a plot of FMR and FNMR that reports system performance at any possible operating point (matching threshold).

It should be highlighted that participants have developed their systems on the basis of a development set with only 10 users, which were them evaluated on 300 sequestered users. Table 2 summarizes the most important characteristics of the best system submitted by each participant, while Table 3 presents the top results achieved across all their submissions (training with first session and testing with remaining three). The results show clear differences between the systems proposed by the participants, whose corresponding EER ranged between 5.32% and 17.90% for the Global EER (EER_G) and 4.72% and 13.66% for the user-dependent EER (EER_U). The large difference between EER_G and EER_U of those systems without score normalization (P1, P2 and P3) suggests the importance of this step, especially when a unique threshold (EER_G) is employed [14][15]. To highlight the impact of the normalization on the performances, the EER_G of the best submission drops from 5.32% to 20.17% when no score normalization is employed. Regarding the differences between the systems it is noticeable the unanimity of features and matchers. The combination of hold time and press-press latency and the classifier based on Manhattan distance were chosen by the two best systems. The largest differences lie in the pre-processing and post-processing techniques applied. Around 10% of the samples have different number of keys pressed (mostly produced by the shift key). The system based on DTW alignment (with feature and score normalization) proposed by the U.S. Army Research Laboratory is the best competing approach. The three P2 self-imposed restrictions have limited their performance for the global EER (EER_G) experiment in comparison with the scenario with user-dependent EER (EER_U). However, we consider there is still room for improvements and performance metrics such as FMR_{100} should be improved before the massive deployment of these technologies.

Table 4 includes the performance (EER_G) obtained using the genuine samples from the second and fourth session (maximum time lapse) for testing and first session for training. The results show a marginal degradation of the performance for all systems, which suggest the stability of the user's performances along the different sessions (more than two months between both sessions). Figure 2 shows the DET curves for all submissions (Fig. 2 Left) and best submissions according to the session evaluated (Fig. 2 Right). The curves show how the submissions made by the participants tend to cluster in different performance ranges and the high robustness against the time lapse (2 months between second and fourth session).

5. Conclusions

This paper presented the first keystroke biometrics ongoing evaluation and the results of an associated offline competition used as baseline. The evaluation, developed on the BEAT platform comprises one of the largest fixed-text keystroke databases available and a fully reproducible benchmark. The performances achieved by the participants are encouraging with a best EER of 5.32%, which could be used as a challenging baseline in further research.

Acknowledgment

A.M. and M. G.-B. are supported by a JdC contract (JCI-2012-12357) and a FPU Fellowship from Spanish MINECO and MCD, respectively. J.M. and J.C. are supported by CAPES and CNPq (grant 304853/2015-1). This work was partially funded by the projects: CogniMetrics (TEC2015-70627-R) from MINECO FEDER and BEAT (FP7-SEC-284989) from EU.

References

[1] A. Peacock, X. Ke, M. Wilkerson, "Typing patterns: A key to user identification", *IEEE Security and Privacy*, 2(5):40–47, 2004.

[2] Y. Zhong and Y. Deng, "A survey on keystroke dynamics biometrics: approaches, advances, and evaluations". Y. Zhong, Y.

[6] https://goo.gl/EQeUBj

Table 2. Summary of the characteristics of the best approaches submitted by the participants.

Participant	Preproc.	Features	Feature norm.	Matcher	Score norm.
P1- Indian Institute of Technology Kharagpur	no	Hold+RP	no	Combined	no
P2 - Federal University of Sergipe	yes	Hold+PP	no	Manhattan	no
P3 - Anonymous participant	no	RP	no	Kendall's tau	no
P4 - U.S. Army Research Laboratory	yes	Hold+PP	yes	Manhattan	yes

Table 3. Final results (best systems) for the KBOC16 offline competition: EER_G (user-independent threshold), EER_U (user-dependent-threshold), FMR_{100}. Training with first session and testing with remaining 3 sessions.

ID	EER_G	EER_U	FMR_{100}
P1	15.73%	11.95%	51.13%
P2	11.82%	7.96%	54.65%
P3	17.90%	13.66%	64.60%
P4	5.32%	4.72%	28.36%

Table 4. Best EER_G for the KBOC16 offline competition according to the session used for testing. Training with first session and testing with second and fourth sessions.

ID	Second Session	Fourth Session
P1	15.28%	16.13%
P2	11.60%	11.96%
P3	17.01%	18.21%
P4	5.09%	5.10%

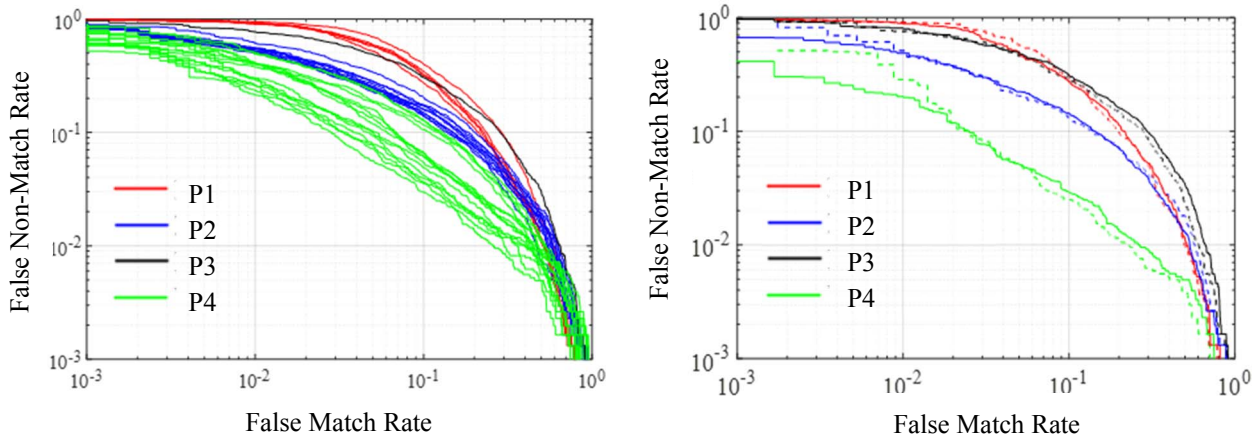

Figure 2. Left: DET curves obtained from all submissions (training with first and testing with remaining 3 sessions); Right: results with different time lapse between enrolment (first session) and testing: testing with second (dashed) and fourth session (solid).

Deng (eds.) *Recent Advances in User Authentication Using Keystroke Dynamics Biometrics*. Science Gate Pub., pp. 1-22, 2015.

[3] 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.

[4] R. Giot, M. El-bed and R. Christophe, "Greyc keystroke: a benchmark for keystroke dynamics biometric systems", *Proc. of IEEE Intl. Conf. on Biometrics: Theory, Applications and Systems*, pp. 1-6, 2009.

[5] C. Loy, W. K. Lai, C. Lim, "Keystroke patterns classification using the ARTMAP-FD neural network", *Proc. of Third Intl. Conf. on Intelligent Information Hiding and Multimedia Signal Processing*, pp. 61–64, 2007.

[6] E. Vural, J. Huang, D. Hou, S. Schuckers, "Shared Research Dataset to Support Development of Keystroke Authentication", *Proc. of Int. Joint Conf. on Biometrics*, pp. 1-8, 2014.

[7] Y. Li, B. Zhang, Y. Cao, S. Zhao, Y. Gao, J. Liu. "Study on the Beihang Keystroke Dynamics Database", *Proc. of Intl. Joint Conf. on Biometrics*, pp. 1-5, 2011.

[8] A. Morales *et al*, "Keystroke Dynamics Recognition based on Personal Data: A Comparative Experimental Evaluation Implementing Reproducible Research", *Proc. of the IEEE Int. Conf. on Biometrics: Theory, Applications and Systems*, pp. 1-6, 2015.

[9] J. V. Monaco, G. Perez, C. C. Tappert, P. Bours, S. Mondal, S. Rajkumar, A. Morales, J. Fierrez, J. Ortega-Garcia, "One-handed Keystroke Biometric Identification Competition", *Proc. IEEE/IAPR Int. Conf. on Biometrics*, pp. 58-64, 2015.

[10] R. Cappelli, M. Ferrara, D. Maltoni, F. Turroni, "Fingerprint Verification Competition at IJCB2011", *Proc. of the IEEE/IAPR Int. Joint Conference on Biometrics*, pp. 1-6, 2011.

[11] S. Marcel, "BEAT biometrics evaluation and testing", *Biometric Technology Today*, pp. 5-7, 2013.

[12] J. Fierrez, *et al.*, "BiosecurID: A Multimodal Biometric Database", *Pattern Analysis and Applications*, 13(2):235-246, 2010.

[13] J. Montalvão, E. O. Freire, M. A. Bezerra Jr., R. Garcia, "Contributions to empirical analysis of keystroke dynamics in passwords", *Pattern Recognition Letters*, 52(15):80-86, 2015.

[14] A. Morales, E. Luna, J. Fierrez, J. Ortega-Garcia, "Score Normalization for Keystroke Dynamics Biometrics", *Proc. of 49th Annual Int. Carnahan Conf. on Security Technology*, pp. 1-6, 2015.

[15] J. Fierrez-Aguilar, J. Ortega-Garcia, J. Gonzalez-Rodriguez, "Target dependent score normalization techniques and their application to signature verification", *IEEE Trans. on Systems, Man & Cybernetics - Part C*, 35(3):418-425, 2005.