Keystroke Biometrics Ongoing Competition

Aythami Morales, Julian Fierrez, Ruben Tolosana, Javier Ortega-Garcia, Javier Galbally,
Marta Gomez-Barrero, Andre Anjos and Sebastien Marcel

Abstract— ***This paper presents the first Keystroke Biometrics Ongoing Competition (KBOC) organized to establish a reproducible baseline in person authentication using keystroke biometrics. The competition has been developed using the BEAT platform and includes one of the largest keystroke databases publicly available based on a fixed text scenario. The database includes genuine and attacker keystroke sequences from 300 users acquired in 4 different sessions distributed in a four month time span. The sequences correspond to the user's name and surname and therefore each user comprises an individual and personal sequence. As baseline for KBOC we report the results of 31 different algorithms evaluated according to accuracy and robustness. The systems have achieved EERs as low as 5.32% and high robustness to multisession variability with accuracy degradation lower than 1% for probes separated by months. The entire database is publicly available at the competition website.***

***Index Terms*—Keystroke, biometrics, authentication, web-biometrics, behavioral recognition, competition, BEAT**

I. INTRODUCTION

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 characteristic used. Behavioral biometrics have boosted the interest of researchers and industry because of their ease of use, transparency and large number of potential applications [1]. Biometric applications have been investigated over the past decades, attracting both academics and practitioners. Biometric recognition systems validate the subject identity by comparing the subject template (pre-stored in a database) with a captured biometric sample [2]. Keystroke biometrics refers to technologies developed for automatic user authentication/identification based on the classification of their typing patterns. 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) and variable performance (*e.g.* human behavior is strongly user-dependent and varies significantly between subjects [3]).

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 online authentication is a real need and platforms such as Coursera use keystroke dynamics to certify the completion of its courses[1]. The number of companies offering keystroke authentication services is large, namely: KeyTrac (www.keytrac.net), Behaviosec (www.behaviosec.com), AuthenWare (www.authenware.com), bioChec (www.biochec.com), ID-Control (www.idcontrol.com), BioValidation (www.biovalidation.com), among other.

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 [1-6] that analyze 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. This scenario usually considers small text sequences as those employed in password authentication services.
- **Free text:** the text used to model the typing behavior and the text used to authenticate do not necessarily match. This scenario is usually related with long text sequences and continuous authentication services.

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 different systems using the same datasets [7-15]. Table 1 summarizes some of the most popular keystroke dynamics public datasets based on fixed text sequences. Even though these benchmarks represent valuable resources, they suffer from two important limitations: (i) The databases available rarely surpass one hundred users. These limited databases decrease the statistical significance of the results and make difficult to establish clear differences between algorithms and methods. (ii) Some of the most popular databases assume that all users share the same password (*e.g.* ".tie5Roanl" and "greyc laboratory" for CMU [9] and GREYC [10] respectively). In real applications, the assumption of different passwords for each user is a more likely scenario. In addition, previous studies suggest that the complexity of the

A. Morales is supported by a JdC contract (JCI-2012-12357) from Spanish MINECO. This work was partially funded by the projects: CogniMetrics (TEC2015-70627-R) from MINECO/FEDER and BEAT (FP7-SEC-284989) from EU.

A. Morales, J. Fierrez, J. Ortega and R. Tolosana are with ATVS Biometric Recognition Group, Universidad Autonoma de Madrid, C\ Francisco Tomás y Valiente, 11, 28049, Madrid, Spain (e-mail: {aythami.morales, julian.fierrez, ruben.tolosana, javier.ortega}@uam.es).

M. Gomez-Barrero is with the da/sec - Biometrics and Internet Security Research Group, Hochschule Darmstadt, Germany (e-mail: marta.gomez-barrero@h-da.de).

J. Galbally is with the European Commission in DG Joint Research Centre, Ispra 21027 (VA), Italy. (email: javier.galbally@jrc.ec.europa.eu).

A. Anjos and S. Marcel are with Centre du Parc, IDIAP Research Institute, Martigny CH-1920, Switzerland (email: andre.anjos@idiap.ch, marcel@idiap.ch).

[1] https://goo.gl/n8BWGR

Table 1. Survey of some of the most popular publicly available databases for fixed text keystroke dynamics recognition.

Year	Database	#users	#samples*	#sessions**	Properties
2009	Killourhy *et al.* [9]	51	50	8	Same password for all users: ".tie5Roanl"
2009	Griot *et al.* [10]	133	12	5	Same password for all users: "greyc laboratory"
2010	Allen *et al.* [11]	104	3-15	1	Three password for all users
2011	Li *et al.* [12]	117	4-16	1	Different password per user
2013	Idrus *et al.* [13]	110	20	1	Five passphrases for all users
2014	Roth *et al.* [14]	51	4	1	Same paragraph for all users
2014	Vural *et al.* [15]	39	20	1	Three password for all users
2015	Antal *et al.* [16]	42	30	2	Same password for all users: ".tie5Roanl"
2015	Morales *et al.* [17]	63	60	2	Different password per user
2016	**KBOC**	**300**	**28**	**4**	**Different password per user**

*Per session; **Per user

password has a large impact on the performance [18]. A performance analysis based on a unique password limits the applicability of the results.

The aforementioned limitations in the performance assessment of keystroke recognition, can be addressed to a large extent through the organization of technological evaluations. These evaluations are usually presented as competitions in which systems provided by different groups can be compared according to common frameworks proposed by third parties. Biometric traits such as fingerprint [19], face [20], speaker [21] or iris [22] have a large tradition of competitions and evaluations with active participation of both the research community and the industry. To the best of our knowledge, there is only one previous keystroke recognition competition: "*One-handed Keystroke Biometric Identification Competition*" [23]. In that 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 the more natural two-handed typing.

Traditional biometric competitions are only operative during a short window of time and this way they only give a static snapshot of the state-of-the-art in a specific research area. One problem with this approach is that it is difficult to encourage researchers to invest their resources and time to participate in these competitions. Without the participation of the main players, the snapshot may be incomplete. In contrast, ongoing competitions provide a dynamic view constantly updated by the community. The FVC-onGoing competition [19] is a successful example with more than 900 participants and more than 4000 algorithms evaluated since 2009 for fingerprint technologies. On the other hand, the absence of platforms to facilitate reproducibility among the keystroke research community has motivated a widespread variety of experimental protocols and evaluation methodologies [5].

As an attempt to move a step forward from the general contexts of keystroke recognition and of biometric competitions described above, the current paper presents the Keystroke Biometrics Ongoing Competition (KBOC). KBOC is the first **fixed text** keystroke competition (in opposition to the free text evaluation described in [23]) that presents two key characteristics that go beyond the usual practice in the field of biometric evaluation campaigns, namely: KBOC is **ongoing** and **reproducible**. This way KBOC tries to address some of the shortcomings currently present in keystroking biometrics, advancing over previous experiences by:

- Proposing the first **ongoing competition on keystroke biometrics**. The competition is carried out over a **fully reproducible** framework based on the BEAT platform[2] [24]. The term reproducible, as it is employed in this work, is defined as a computational experiment that can be repeated using the same data and tools. The main aim of the competition is to provide a new benchmark that guarantees a fair comparison between keystroke recognition algorithms using the same experimental framework.
- Reporting a **large performance evaluation** of keystroke dynamics technologies including 31 keystroke recognition systems from 4 different research laboratories. The evaluation is performed on the basis of performance and robustness of the different approaches.
- Disclosing a **public database** 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).

The rest of the paper is organized as follows. Section 2 introduces the ongoing evaluation tool developed for KBOC. Section 3 describes the database and evaluation protocols. Section 4 sketches the best systems submitted so far by participants to KBOC (this initial stage of the competition will be referred to as KBOC Baseline). Section 5 reports the experiments and results of KBOC baseline. Section 6 summarizes the conclusions.

II. KBOC Infrastructure

KBOC exploits the potential of the BEAT platform, which was created under the FP7 EU BEAT project to promote reproducible research in biometrics [24]. The BEAT platform is a European computing e-infrastructure for Open Science that proposes a solution for open access, scientific information sharing and re-use of data and source code while protecting privacy and confidentiality. The platform is a web-application allowing experimentation and testing in pattern recognition.

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

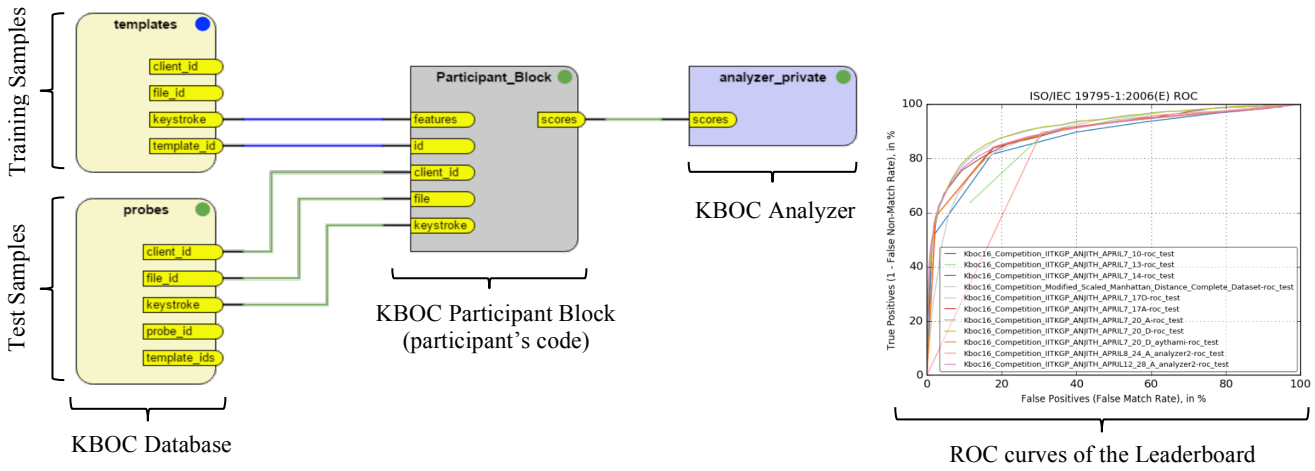
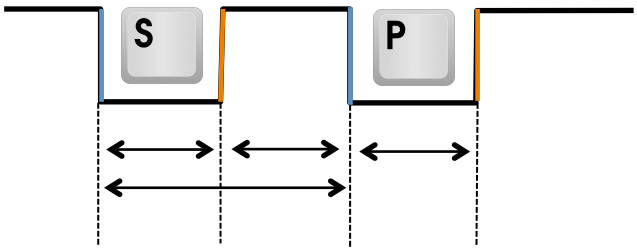


Figure 1. Toolchain of KBOC developed on BEAT (https://goo.gl/8DJQN7).

KBOC provides the data and modules necessary to run the evaluation and the BEAT platform ensures that the system is correctly executed, also producing the results. Different algorithms and systems can be easily compared. The platform also provides an attestation mechanism that guarantees that a certain result has been produced using the BEAT platform, based on some database, protocol and algorithms. This attestation mechanism produces a link that can be included in a report (e.g., scientific papers, technical documents or certifications) so that readers can go to the platform and check the authenticity of the results, even being able to replicate the experiments. There is no limit regarding the number of systems to evaluate, and the results are automatically provided to the participants on the platform (*i.e.*, the performance of the systems is available in real time). KBOC, as part of the BEAT platform, is a web application divided into:

- Toolchain: determines the data flow of the experiments. The toolchain is defined by a block diagram (see Figure 1) including datasets and algorithms. The blocks of KBOC toolchain are: (i) Database: participants cannot access directly the data but they can use it in the experiments. The dataset blocks (templates and probes) define the experimental protocol and they cannot be modified by the users. The platform automatically provides the training samples (labeled data) and test samples (unlabeled data) to the Participant Block. (ii) Participant Block: includes the algorithm to compare keystroke sequences. Participants can modify the code of this block including their keystroke recognition algorithms. The inputs are the samples of the database (training and test samples), and the output are the similarity scores. (iii) Analyzer: this block is the output of the platform. Its tasks include analyzing the scores produced by the participant block and reporting performance according to some standard metrics. Participants can use the analyzer but cannot access its code. This way it is guaranteed that all algorithms are evaluated according to the same parameters.
- Dataformats: describe the information transmitted between blocks of the toolchain. They specify the format of inputs and outputs of the algorithms and databases. KBOC includes a specific dataformat (called kboc16_keystroke[3]) to define the timestamp sequences associated to each sample. The data format includes both the timestamp and the key pressed.
- Leaderboard: represents the experimental results. KBOC is an ongoing competition, therefore the results will be automatically updated with new submissions[4].



It should be noted that participating in the ongoing evaluation and using of the platform do not imply the publication of the code. Confidentiality is a priority and is granted in all cases. 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. Participants retain all access rights to their code. They can keep it private, share it with other specific users, or make it public so that other platform users can benefit and reuse it. These access rights can be different for different parts of the code (e.g., the participant can decide to make public a specific segmentation module but not the matcher).

KBOC is now active and several baseline experiments are available at the BEAT platform[5]. For further

[3] https://goo.gl/lwyBVb
[4] https://goo.gl/EQeUBj
[5] https://goo.gl/VsKgVM

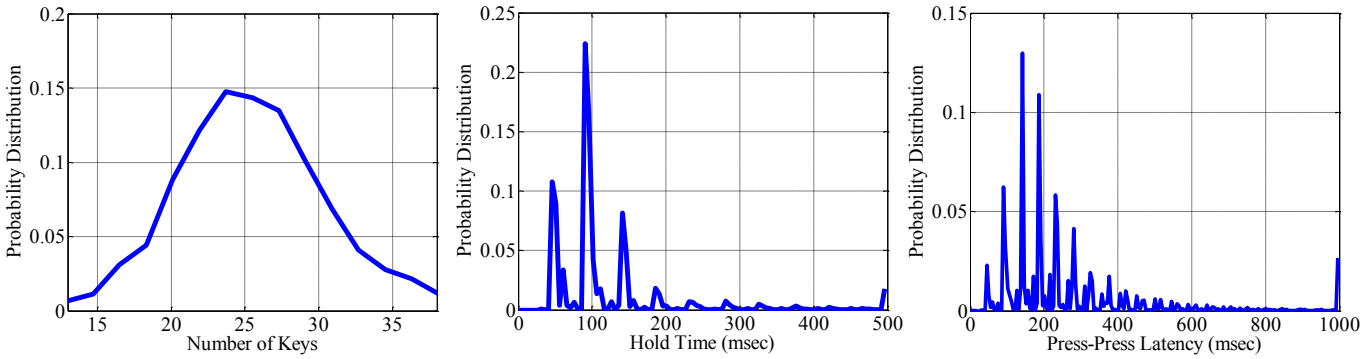


Table 2. Summary of the main statistics of the database proposed for the competition.

Characteristics	#
Number of users (Testing Set)	300
Age distribution	
18–25	42%
25–35	22%
35–45	16%
>45	20%
Handedness	
Righthanded	93%
Lefthanded	7%
Number of users (Development Set)	10
Number of sessions	4
Samples per session*	4-7
*Genuine**	2-4
*Impostor**	2-3
Training samples per user	4
Genuine comparisons per user*	8-12
Impostor comparisons per user*	8-12
Total genuine comparisons	3028
Total impostor comparisons	2972
Clock resolution	~40msec
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.

operational/logistic information on how to participate in the competition please visit the official competition website[6].

III. Dataset and Evaluation Protocols

The dataset proposed for the competition is part of the BiosecurID multimodal database [25] 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. 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. Therefore, the discriminative power of these sequences is larger than other random free text scenarios.

The BiosecurID multimodal 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 sequences provided to the participants include the time intervals between consecutive key events (press and release) and the ANSI code associated to the key pressed. Figure 2 shows the timestamps and three of the most popular features used in keystroke dynamics: Hold Time ($t_i^r - t_i^p$), Press-Press Latency ($t_{i+1}^p - t_i^p$) and Release-Press Latency ($t_{i+1}^p - t_i^r$). The main statistics of the dataset proposed for the competition are summarized in Table 2 and probability distributions of some key features are showed in Figure 3. The statistics show that most sequences have length ranging from 13 to 38 characters (see Figure 3 left). Regarding two of the most

[6] https://sites.google.com/site/btas16kboc/

Table 3. 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 - Universidade Federal de Sergipe	yes	Hold+PP	yes	Manhattan	no
P3 - Anonymous participant	no	RP	no	Kendall's tau	no
P4 - U.S. Army Research Laboratory	yes	Hold+PP	yes	Manhattan	yes

popular characteristics on keystroke dynamics, the values of Hold Time (difference between timestamps of press and release events of the same key) are distributed around 3 values (as can be seen in Figure 3 center), while Press-Press Latency (difference between timetimestamps of press and press events of consecutive keys) are distributed around more than 8 values (see Figure 3 right). As it can be seen in Figure 3, the clock resolution is approximately 40 msec. The clock resolution defines precision of the timestamps (i.e. the maximum difference between the real timestamp and the measured timestamps is ±40 msec). The use of external reference clocks can be used to increase the resolution of keystroke latencies [26] and improve the performance of recognition systems.

The experimental protocol is 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 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 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 only difference between both modes is that for the offline competition was organized as part of the The IEEE Eighth International Conference on Biometrics: Theory, Applications, and Systems (BTAS 2016) and therefore a deadline was set for the submission of algorithms. The performance of the offline evaluation (detailed in section 5) will be used as baseline for the ongoing competition. The complete dataset is available at KBOC website[7].

IV. KBOC Baseline: Systems

In order to start up KBOC, a traditional offline competition was first proposed to serve as KBOC Baseline [27]. The training set and test set (described in section 3) were available for all the participants. 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.

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 of the registered institutions finally submitted their systems for a total of 31 evaluated systems.

The systems evaluated include the most popular machine learning algorithms (Neural Networks, Support Vector Machines, Decision Trees) as well as basic distances (Euclidean, Manhattan, Mahalanobis) popular in keystroke dynamics literature. Different strategies were proposed to normalize features and scores. The best system of each of the three non-anonymous participants is briefly below. Table 3 summarizes the most important characteristics of the best system submitted by each participant.

A. U.S. Army Research Laboratory (ARL)

The main characteristics of the best system (number 6 of 15) submitted by ARL team is summarized as: (i) The features used are Hold Time and Press-Press Latency. (ii) ARL team proposes an element-wise semantic alignment (see [28] for details) between the target sequence (minimum length sequence in the training data) and the query sequence. A modified Dynamic Time Warping (DTW) algorithm is used to match multiple minimum length sequences (misaligned samples). (iii) The features (Hold Time and Press-Press Latency) are normalized according to the following equation:

$$\hat{f}_i^j = \max\left(0, \min\left(1, \frac{f_i^j - \lfloor f \rfloor}{\lceil f \rceil - \lfloor f \rfloor}\right)\right) \quad (1)$$

where $\lfloor f \rfloor$ and $\lceil f \rceil$ are the lower and upper bounds respectively defines as $\lfloor f \rfloor = \mu - \sigma$ and $\lceil f \rceil = \mu + \sigma$. The mean μ and standard deviation σ are calculated from all the training samples. (iv) The distance between a query sequence and the training set is calculated using the Manhattan distance as:

$$d = \sum_{i=1}^{M} \left| \hat{f}_i^j - \bar{g}_i^j \right| \quad (2)$$

were $\bar{\mathbf{g}}^j$ is the mean training vector. (v) Finally, the distances from query samples to each claimed identity are then normalized similarly to Eq. (1) to within $\pm 2\sigma$ of the mean, with distances outside that range clipped to $[0,1]$. In that case,

[7] https://sites.google.com/site/btas16kboc/home

the lower and upper bounds are calculate as $\lfloor d^j \rfloor = \mu^j - 2\sigma^j$ and $\lceil d^j \rceil = \mu^j + 2\sigma^j$ with μ^j and σ^j the mean and standard deviation of the user j.

The code of all 15 systems submitted by ARL team to KBOC are available at[8]. See [28] for a detailed description of all systems.

B. Universidade Federal de Sergipe (UFS)

The main characteristics of the best system (number 7 of 10) submitted by UFS team is summarized as: (i) The features used are Hold Time and Press-Press Latency. (ii) UFS team does not propose any alignment procedure but includes a shuffling procedure [29] to mitigate it. In case of inconsistent length sequences, the minimum one is compared to each sub segment of the longer one and the minimum distance is kept. (iii) As in [18], Press-Press Latency (PP) and Hold Time (H) intervals are normalized 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:

$$\hat{f}_i^j = \left(1 + \exp\left(-\frac{1.7(\log_e(f_i^j) - \mu)}{\sigma}\right)\right)^{-1} \quad (3)$$

where f_i^j stands for a time interval i (in seconds) of user j. (iv) The distance between a query sequence and the training set is calculated using a modified Manhattan distance as:

$$d = \frac{1}{4M}\sum_{i=1}^{M}\left|\hat{f}_{h,i}^j - \bar{g}_{h,i}^j\right| + \frac{3}{4(M-1)}\sum_{i=1}^{M-1}\left|\hat{f}_{l,i}^j - \bar{g}_{l,i}^j\right| \quad (4)$$

were $\hat{\mathbf{f}}_h^j$, $\hat{\mathbf{f}}_l^j$ are the normalized Hold Time and Press-Press Latency features respectively and $\bar{\mathbf{g}}_h^j$, $\bar{\mathbf{g}}_l^j$ are the gallery features. (v) Finally, the UFS team proposes a strategy to update the training set every time a query sample obtain a score lower than 0.14.

C. Indian Institute of Technology Kharagpur (IITK)

The main characteristics of the best system (number 5 of 5) submitted by IITK team is summarized as: (i) The features used are Hold Time and Release-Press Latency. (ii) The distance between a query sequence and the training set is calculated using two distance metrics based on mean and median. The distance measures were computed as:

$$\boldsymbol{\Delta}_i^{j,k} = \left|g_i^{j,k} - f_i^j\right|, \qquad k = 1, \ldots 4 \text{ and } i \in [1, \ldots, M] \quad (5)$$

$$\boldsymbol{\lambda}_i^j = \min_{k\in[1,\ldots,4]} \delta_i^{j,k}, \qquad i \in [1, \ldots, M] \quad (6)$$

where $\delta_i^{j,k}$ is an element of matrix $\boldsymbol{\Delta}_i^{j,k}$ and the final distance was obtained as:

$$d = \mathrm{mean}(\boldsymbol{\lambda}) + \mathrm{median}(\boldsymbol{\lambda}) \quad (7)$$

V. KBOC Baseline: Experiments

The participants were allowed to submit up to 15 different systems before the deadline. The test samples remained sequestered (*i.e.*, participants did not know whether they were genuine or impostors samples). In addition, a small development set (10 users with labeled samples) and baseline algorithms were provided to the participants following the instruction given in the competition website and upon the signing of an agreement in order to access these personal data. 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 ($\mathrm{EER_G}$): unique EER calculated using all genuine and impostor scores and only one decision threshold for all users. 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.
- User-dependent Equal Error Rate ($\mathrm{EER_U}$): the EER is calculated independently for each of the 300 subjects (300 different decision thresholds). $\mathrm{EER_U}$ is the average individual EER from all subjects. This EER is common in the keystroke dynamics literature [4],[9],[10].
- Detection-Error Tradeoff (DET) curve: a plot of FMR and FNMR that reports system performance at any possible operating point (decision threshold).

A. Performance Evaluation

It should be highlighted that participants have developed their systems on the basis of a development set with only 10 users, which were then evaluated on 300 sequestered users. Table 4 presents the 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 ($\mathrm{EER_G}$) and 4.72% and 13.66% for the user-dependent EER ($\mathrm{EER_U}$). The large difference between $\mathrm{EER_G}$ and $\mathrm{EER_U}$ of those systems without score normalization (P1, P2 and P3) suggests the importance of this step, especially when a unique threshold ($\mathrm{EER_G}$) is employed [3][30]. To highlight the impact of the normalization on the performances, system 5 of P4 was evaluated (after the competition and once the results were published) without the score normalization. The $\mathrm{EER_G}$ achieved by this system drops from 5.32% to 20.17% when no score normalization is employed.

Figure 4 left shows the DET curves for all submissions. The curves show how the submissions made by the participants tend to cluster into different performance ranges. Regarding the differences between the systems (see Table 3) 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 used by the two best participants (P2 and P4). 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 to reduce the gap between the global EER ($\mathrm{EER_G}$) and the user-dependent EER ($\mathrm{EER_U}$) that results on improved performances. Further sections will analyze the results depending on different factors.

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

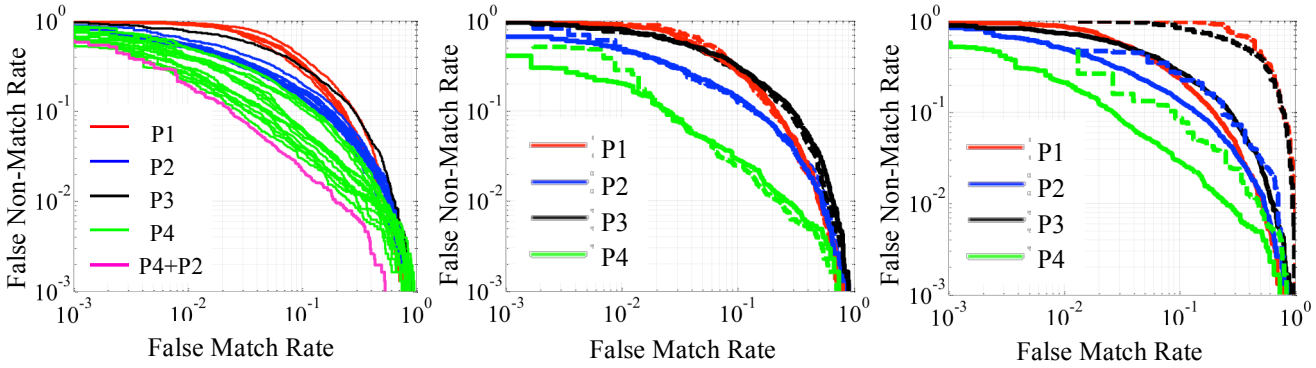


Figure 4. Left: DET curves obtained from all submissions (training with first and testing with the remaining 3 sessions) and the combination of the best system of P2 with the best system of P4 (P4 and P2 are the respectively the two best participants). Center: results of the best systems with different time span between enrolment (first session) and testing: testing with second (dashed) and fourth session (solid). Right: results of the best systems with aligned samples (solid) and misaligned samples (dashed).

Table 4. Final results (all systems) in KBOC baseline: EER_G (user-independent-threshold) / EER_U (user-dependent threshold). Training with first session and testing with the 3 remaining sessions. P1 to P4 as in Table 3. Rows indicate different systems submitted by the same participant (best participant codes P1 to P4 are available in Table 3).

#	P1	P2	P3	P4
1	19.33/11.49	12.90/9.61	**17.90/14.36**	7.82/7.32
2	16.82/12.33	11.85/7.70		6.46/6.03
3	16.52/12.01	12.12/9.28		7.32/6.67
4	16.47/12.11	13.48/10.38		7.35/6.37
5	**15.73/11.26**	12.25/8.99		8.02/7.66
6		13.92/10.42		**5.32/4.62**
7		**11.82/8.81**		7.95/7.27
8		13.03/9.90		8.08/7.87
9		13.03/9.52		5.68/5.46
10		14.66/11.66		5.91/5.36
11				10.35/6.67
12				10.89/6.37
13				11.20/7.66
14				11.23/7.87
15				6.26/5.65

B. Robustness Against Time-Lapse

As a behavioral trait, the robustness of keystroke biometrics to increasing time between enrollment and testing is an important factor to consider [31]. The database employed allows to analyze the performance for different intervals between enrollment and testing: few weeks (session 1 vs session 2) and few months (session 1 vs session 4). Table 5 includes the performance (EER_G) obtained using the genuine samples from the second or fourth session for testing and the samples from the first session for training. The results show a significant robustness of all systems to this time-lapse (slightly over 2 months), presenting a small performance drop always under 10%. Even systems with moderate performance show high stability of the genuine scores for the different sessions. These results can be seen at Figure 4 center and Figure 5, where it is possible to observe that genuine scores from different sessions show almost identical distributions. Note that, as specified in section 3, the keystroke sequences used in this work are very familiar sequences, namely: name and surname. These results suggest that keystroke dynamics based on such information remain consistent even for acquisitions separated by months.

Table 5. EER_G for the best system of each participant according to the session used for testing. Training with first session and testing with second and fourth sessions. The last row indicates the drop of performance between sessions.

Session	P1	P2	P3	P4
Second	15.28%	11.60%	17.01%	5.09%
Fourth	16.13%	11.96%	18.21%	5.10%
Difference	↑5%	↑3%	↑7%	↑<1%

C. Robustness Against Key Sequence Misalignment

As it was described in section 3, 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. Table 6 and Figure 4 right show the performances obtained by the best systems for the aligned samples (sequences with exactly the same keys) and the misaligned samples (samples with different length and therefore different keys). In general, there is a significant drop of performance between both sets that can be more than 300%. The strategy based on DTW alignment adopted by the U.S. Army Research Laboratory shows the best performance in both types of samples. How to deal with these misaligned samples is still an open challenge to be explored by the research community [32].

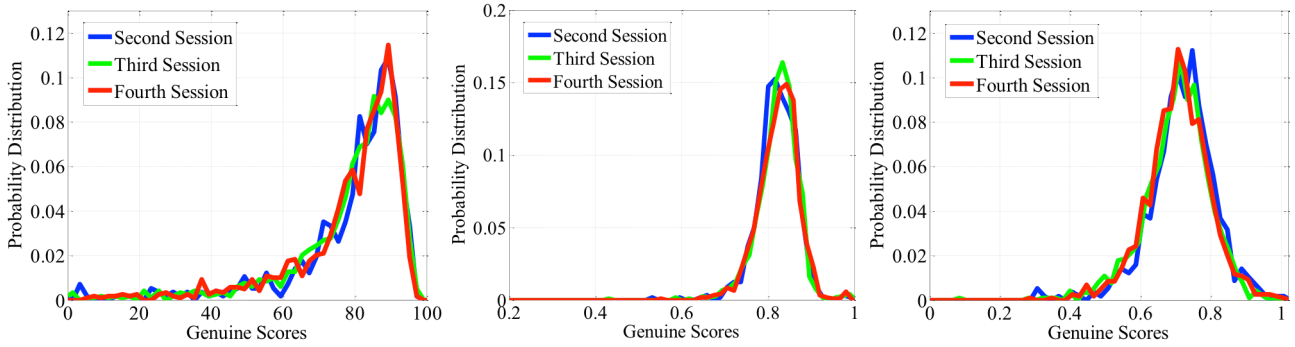
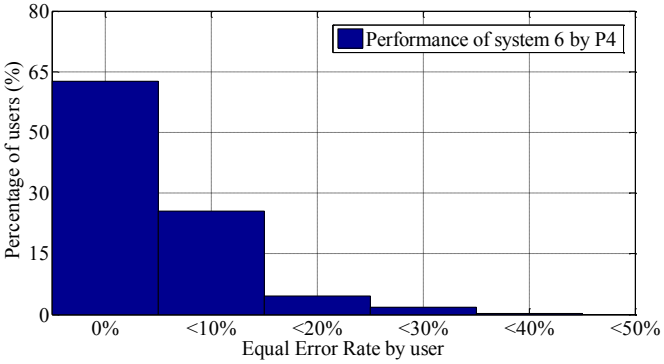
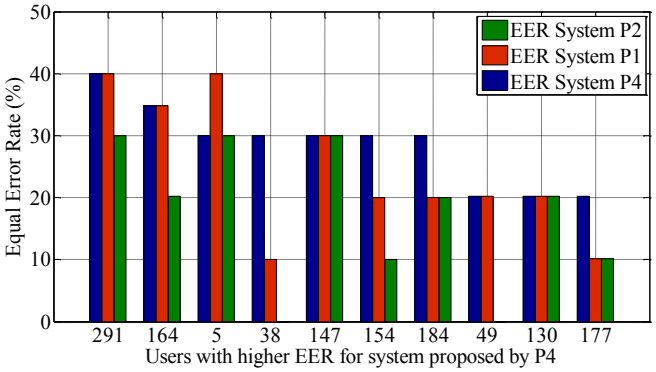


Figure 5. Genuine score distribution according to the different sessions used for testing: P1 system 5 (Left), P2 system 7 (Center) and P4 system 6 (Right).

Table 6. EER_G for the best system of each participant according to the nature of the samples used for testing. The last row indicates the drop of performance between aligned and misaligned samples.

Samples	P1	P2	P3	P4
Aligned	14.75%	11.60%	17.11%	5.21%
Misaligned	48.63%	18.21%	48.68%	9.20%
Difference	↑329%	↑157%	↑284%	↑176%





D. User-dependent Performance

The performance of keystroke dynamics is strongly user-dependent [33]. As an example, Figure 6 shows the histogram (in terms of probability distribution) of the EER of the best system evaluated in the competition (system 6 by P4) obtained independently for each of the 300 users. The results show a large margin between performances of different users with substantial percentage of users with differences of EER up to 20%. How to improve the performance of the worst users is an open challenge in keystroke dynamics. One possibility is to explore the complementarity between algorithms. Figure 7 shows the EER of the users with worst performance using the best system submitted by P4 and the performance obtained for the same users using the systems submitted by P1 and P2. The systems submitted by P1 and P2 show a worse overall performance (see Table 3) than those submitted by P4. However, the results shown in Figure 7 suggest there is a potential complementarity between systems, as P1 and P2 tend to give better results than P4 for these problematic users.

In order to evaluate the complementarity between the different systems we have combined them at score level by a weighted sum [34]. The results (Figure 4 left) suggest certain level of complementarity and the combination of the best systems from P2 and P4 shows the best performance of all systems.

VI. Conclusions

This paper presented the Keystroke Biometric Ongoing Competition (KBOC) and the results of an associated offline competition used as KBOC baseline. The evaluation, developed on the BEAT platform, comprises one of the largest fixed text keystroke databases available. The main characteristics of KBOC can be summarized as: (i) Large evaluation database with 300 users and 7200 keystroke sequences including different passwords for each user. (ii) Multisession database with 4 different sessions across 4 months. Enrollment using samples from the first session and testing with the 3 remaining sessions. (iii) Baseline for a total of 31 keystroke dynamics systems considering both global EER (EER_G) and user-dependent EER (EER_U). (iv) Ongoing tool implementing reproducible research now publicly available based on the BEAT platform.

The experiments reported as KBOC Baseline comparing 31 systems from 4 participants have permitted us to obtain the following new insights to the problem of biometric person recognition based on keystroke dynamics. In first place, it is possible to obtain competitive performances with EER under 6% even in challenging conditions with a small development set of 10 users and test set with 300 users. Secondly, the alignment of sequences with different lengths and the score normalization have showed large potential to improve the systems accuracy. Thirdly, the robustness to a time lapse of two months is remarkable even for those systems with the poorest results. Finally, the performance of keystroke dynamics is highly user-dependent. How to adapt algorithms to the different user behaviors, including synthetic samples [35], remains an open research field.

These observations motivate us to conduct further research in: (i) Score normalization techniques to improve the performance of systems based on unique classification thresholds. (ii) Exploit and explore user-dependencies in order to adapt the algorithms to the variable behavior of users. (iii) New research on alignment strategies to reduce the severe drop of accuracy due to typos.

REFERENCES

[1] A. Peacock, X. Ke and M. Wilkerson, “Typing patterns: A key to user identification”, *IEEE Security and Privacy*, vol. 2, no. 5, pp. 40–47, 2004.

[2] S. Prabhakar, S. Pankanti, and A. K. Jain, “Biometric Recognition: Security and Privacy Concerns”, IEEE Security & Privacy, pp. 33-42, 2003.

[3] J. Fierrez-Aguilar, J. Ortega-Garcia and J. Gonzalez-Rodriguez, "Target dependent score normalization techniques and their application to signature verification", *IEEE Trans. on Systems, Man & Cybernetics - Part C*, vol. 35, no. 3, pp. 418-425, 2005.

[4] Y. Zhong and 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.

[5] 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.

[6] 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.

[7] S. Banerjee and D. Woodard, “Biometric authentication and identification using keystroke dynamics: a survey”, *Journal of Pattern Recognition Research*, 7 (1), pp. 116–139, 2012.

[8] P. S. Teh, A. B. J. Teoh and S. Yue, “A survey of keystroke dynamics biometrics”, *The Scientific World Journal*, pp. 1–24, 2013.

[9] 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.

[10] 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.

[11] J. D. Allen, “An analysis of pressure-based keystroke dynamics algorithms”, Master’s thesis, Southern Methodist University, Texas, 2010.

[12] Y. Li, B. Zhang, Y. Cao, S. Zhao, Y. Gao and J. Liu. “Study on the Beihang Keystroke Dynamics Database”, *Proc. of Int. Joint Conf. on Biometrics*, Washington DC, USA, pp. 1-5, 2011.

[13] S.Z.S. Idrus, E. Cherrier, C. Rosenberger and P. Bours, “Soft biometrics database: a benchmark for keystroke dynamics biometric systems”, *Proc. of 2013 Int. Conf. of the Biometrics Special Interest Group*, pp. 1–8, 2013.

[14] J. Roth, X. Liu, and D. Metaxas, “On continuous user authentication via typing behavior,” *IEEE Trans. Image Processing*, vol. 23, no. 10, pp. 4611–4614, 2014.

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

[16] M. Antal, L. Z. Szabó, and I. László, “Keystroke Dynamics on Android Platform”, *Procedia Technology*, vol. 19, pp. 820–826, 2015.

[17] A. Morales, M. Falanga, J. Fierrez, C. Sansone and J. Ortega-Garcia, "Keystroke Dynamics Recognition based on Personal Data: A Comparative Experimental Evaluation Implementing Reproducible Research", *Proc. of the IEEE Seventh Int. Conf. on Biometrics: Theory, Applications and Systems*, Arlington, Virginia, USA, pp. 1-6, 2015.

[18] 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.

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

[20] G. B. Huang and E. Learned-Miller, “Labeled faces in the wild: Updates and new reporting procedures”, *Report from University of Massachusetts*, Amherst, UM-CS-2014-003, 2014.

[21] C. Greenberg, D. Banse, G. Doddington, D. Garcia-Romero, J. Godfrey, T. Kinnunen, A. Martin, A. McCree, M. Przybocki, and D. Reynolds, “The NIST 2014 speaker recognition i-vector machine learning challenge,” *in Odyssey 2014: The Speaker and Language Recognition Workshop*, Joensuu, Finland, 2014.

[22] P. Jonathon Phillips, Patrick J. Flynn, J. Ross Beveridge, W. Todd Scruggs, Alice J. O’Toole, David Bolme, Kevin W. Bowyer, Bruce A. Draper, Geof H. Givens, Yui Man Lui, Hassan Sahibzada, Joseph A. Scallan III and S. Weimer, “Overview of the multiple biometrics grand challenge”, *Advances in Biometrics*, LNCS. vol. 5558, pp. 705–714. Springer Berlin Heidelberg, 2009.

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

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

[25] J. Fierrez, J. Galbally, J. Ortega-Garcia, M. R. Freire, F. Alonso-Fernandez, D. Ramos, D. T. Toledano, J. Gonzalez-Rodriguez, J. A. Siguenza, J. Garrido-Salas, E. Anguiano, G. Gonzalez-de-Rivera, R. Ribalda, M. Faundez-Zanuy, J. A. Ortega, V. Cardeñoso-Payo, A. Viloria, C. E. Vivaracho, Q. I. Moro, J. J. Igarza, J. Sanchez, I. Hernaez, C. Orrite-Uruñuela, F. Martinez-Contreras, J. J. Gracia-Roche, "BiosecurID: A Multimodal Biometric Database", *Pattern Analysis and Applications*, vol. 13, no. 2, pp. 235-246, 2010.

[26] K.S. Killourhy and R. A. Maxion, “The effect of clock resolution on keystroke dynamics”. In: R. Lippmann, E. Kirda, A. Trachtenberg (eds.) *Lecture notes in computer science*, vol. 5230. Springer, pp. 331-350, 2008.

[27] A. Morales, J. Fierrez, M. Gomez-Barrero, J. Ortega-Garcia, R. Daza, J. V. Monaco, J. Montalvão, J. Canuto and A. George, "KBOC: Keystroke Biometrics OnGoing Competition", *Proc. 8th IEEE Int. Conf. on Biometrics: Theory, Applications, and Systems*, Buffalo, USA, pp. 1-6, 2016.
[28] J. V Monaco, "Robust Keystroke Biometric Anomaly Detection", *arXiv preprint*, pp. 1-7, 2016.
[29] S. Bleha, "Recognition systems based on keystroke dynamics", Ph.D. thesis, Univ. Missouri, Columbia, 1998.
[30] A. Morales, E. Luna, J. Fierrez and J. Ortega-Garcia, "Score Normalization for Keystroke Dynamics Biometrics", *Proc. 49th Annual Int. Carnahan Conf. on Security Technology*, pp. 1-6, Taipei, Taiwan, 2015.
[31] J. Galbally, M. Martinez-Diaz and J. Fierrez, "Aging in Biometrics: An Experimental Analysis on On-Line Signature", *PLOS ONE*, nol. 8, no. 7, pp. e69897, 2013.
[32] P. Bours and V. Komanpally, "Performance of keystroke dynamics when allowing typing corrections", *Proc. of IEEE International Workshop on Biometrics and Forensics*, Valletta, Malta, pp. 1-6, 2014.
[33] A. Morales, J. Fierrez and 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, Zurich, Switzerland, 2014.
[34] J. Fierrez, D. Garcia-Romero, J. Ortega-Garcia and J. Gonzalez-Rodriguez, "Adapted user-dependent multimodal biometric authentication exploiting general information", *Pattern Recognition Letters*, vol. 26, no. 16, pp. 2628-2639, 2005.
[35] D. Stefan, X. Shu and D. Yao, "Robustness of keystroke-dynamics based biometrics against synthetic forgeries", *Computers & Security*, vol. 31, pp. 109-121, 2012.