COMPUTERS & SECURITY 55 (2015) 46–61

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/cose

Computers & Security

A review on the public benchmark databases for static keystroke dynamics

Romain Giot [a], *Bernadette Dorizzi* [b], *Christophe Rosenberger* [c,*]

[a] *University of Bordeaux – LaBRI, 351 cours de la Libération F-33405, Talence cedex, France*
[b] *Institut Mines Télécom/Télécom SudParis (IMT/TSP), 9 rue Charles Fourier, 91011 EVRY Cedex, France*
[c] *GREYC, ENSICAEN – UNICAEN – CNRS, 6 boulevard Maréchal juin, 14000 Caen, France*

ARTICLE INFO

Article history:
Received 28 July 2014
Received in revised form 11 May 2015
Accepted 26 June 2015
Available online 4 August 2015

Keywords:
Keystroke dynamics
Benchmark algorithms
Biometric performance evaluation
Online and offline computation

ABSTRACT

Keystroke dynamics allows to authenticate individuals through their way of typing their password or a free text on a keyboard. In general, in biometrics, a novel algorithm is validated through a comparison to the state of the art one's using some datasets in an offline way. Several benchmark datasets for keystroke dynamics have been proposed in the literature. They differ in many ways and their intrinsic properties influence the performance of the algorithms under evaluation. In this work, we (a) provide a literature review on existing benchmark datasets of keystroke dynamics; (b) present several criteria and tests in order to characterize them; (c) and apply these criteria on these available public benchmark datasets. The review analysis shows a great disparity in the acquisition protocol, the population involved, the complexity of the passwords, or the expected performance (there is a relative difference of 76% between the EER on the worst and best performing datasets with the same authentication method).

1. Introduction

Keystroke dynamics (Giot et al., 2011) is a behavioural biometric modality which allows the authentication of individuals through their way of typing their password on a keyboard. It is a behavioural biometrics which presents the advantage of not requiring additional sensor than the keyboard at hand and which allows authentication through time (on-line authentication). It is probably the cheapest biometric modality available of a personal computer.

Biometric systems are validated, thanks to datasets that are collected for this purpose and can be qualified in function of the problem that one wants to solve. For instance, if one wants to benchmark algorithms for face recognition, the algorithms may give very different performance on a dataset which contains no variability (visage is frontal with no expression and no illumination variation) and on other dataset which contains some of these variabilities. The choice of the dataset is therefore guided by the precise problem the algorithm aims at solving. A good evaluation framework corresponds to the association of a reference system, a dataset and an associated evaluation protocol which fixes the way the dataset is used. This way, it is possible to compare different systems between them without providing any bias and to report comparative results using some error measures which very often in biometrics corresponds to EER and ROC curves (Mansfield and Wayman, 2002; Poh and Bengio, 2006). Such protocols have been proposed for most of the biometric modalities as described in Jain et al. (2009). Keystroke dynamics is a recent modality for which no such framework has been yet proposed. One of the aims of the present work is to give some report of the

* *Corresponding author.* Tel.: +33 231538135.
E-mail address: christophe.rosenberger@ensicaen.fr (C. Rosenberger).
http://dx.doi.org/10.1016/j.cose.2015.06.008
0167-4048/

existing situation in terms of publicly available datasets and to suggest some criteria that could guide the construction of new datasets helpful for the research community.

Although the authentication is done in real time (*i.e.*, online) in a real world system, scientists working on keystroke dynamics do not analyse the performance of their system online (*i.e*, by asking users to authenticate themselves in real time and to impersonate other users). Indeed, they work in an offline way by using samples previously collected, probably by other researchers, and stored in a benchmark dataset. Thus, data collection and algorithms evaluation are often two separated tasks. The evaluation of algorithms is eased, thanks to this offline procedure. If scientists share the same common benchmark, they can fairly compare their algorithms by considering error results and time consumption. Experiments become easily reproducible, which is an important criterion in scientific studies (although datasets could contain errors). Data collection is a tedious and time consuming task, which can explain why there are only few benchmark datasets of high quality. It is important to characterize the datasets in order to be able to easily compare them, and to choose the adequate one depending on the study.

A keystroke dynamic (KD) system (KDS) is composed of two main modules: the enrollment and the verification modules. Each user must enroll himself in the KDS which computes a biometric reference given multiple samples (*i.e.*, several inputs of the password) acquired during the enrollment step. For each input, a sequence of timing information is captured (*i.e.*, time when each key is pressed or released) from which some features are extracted (*i.e.*, latencies and durations) and used to learn the model which characterizes each user. During a verification request, the claimant types his/her password. The system extracts the features and compares them to the biometric reference of the claimant. If the obtained distance is below a threshold, the user is accepted, otherwise he/she is rejected. An optional module can be used to automatically update the model of the user (Giot et al., 2011). This can be important, as the KD data are not permanent and evolves with time (Giot et al., 2011; Kang et al., 2007). First works on KD have been done in the eighties (Gaines et al., 1980), although the idea of using a keyboard to automatically identify individuals has first been presented in 1975 (Spillane, 1975). In the preliminary report of Gaines et al. (1980), seven secretaries typed several paragraphs of text and researchers showed that it is possible to differentiate users with their typing patterns. Since then, several studies have been done, allowing to decrease the quantity of information needed to build the biometric reference, while improving the performances (Giot et al., 2011; Lee and Cho, 2007; Monrose and Rubin, 2000; Revett et al., 2007; Umphress and Williams, 1985). However, most studies are not comparable because they use different datasets or protocols (Giot et al., 2011; Killourhy and Maxion, 2011).

In this paper, we present a review of existing publicly accessible benchmark datasets for keystroke dynamics. We also propose a methodology to characterize these KD benchmark datasets (KDB). Although it would be fruitful to rank them according to some criteria, it is actually difficult to give them a score because such score cannot be generic, as it would depend on the kind of study we want to do, and would be based on too many subjective points. Consequently, in this work, we aim at considering all the interesting points to analyse in order to qualify a KDB; using this information, the scientist will be able to choose the best dataset adapted to his particular experiment. Alternatively, it can also help him to create a new dataset of higher quality than the existing ones and may be more suitable to the problem he wants to tackle with. In this work, we focus only on static password KD authentication systems (*i.e.*, each user is authenticated, thanks to the typing of an expected password, and not free text). The originality of this work is (a) the proposition of a complete set of criteria for characterizing different datasets, which is lacking in this field, and (b) the analysis of all the existing public KDB with respect to these criteria. We think this work is important because it is known that KD studies are not fair as (i) acquisition protocols are different between studies (Giot et al., 2011); (ii) there is not always a comparative study (Killourhy and Maxion, 2011) when authors propose new algorithms; and (iii) there are not always a valuable statistical evaluation (Killourhy and Maxion, 2011). Our work helps to solve the two first problems, while methods presented in (Schuckers, 2010) can solve the third one.

The paper is organized as follows. Section 2 presents a review of existing benchmark datasets for keystroke dynamics. Section 3 presents the various important elements to consider when characterizing KDB. Section 4 realizes a comparative study of public KDB by using all the criteria defined in the previous sections. Section 5 concludes this paper.

2. Existing benchmark datasets for keystroke dynamics

Contrary to other biometric modalities, there are only a few public datasets for KD. Teh et al. (2013) list three public KDB. However, we found 4 additional ones, but it still stays a low number, in comparison to face recognition, for example. These datasets are detailed in section 4 considering all the important criteria defined in section 3. Note that some of these datasets can be composed of several sub-datasets.

2.1. *GREYC*

Giot et al. propose the most important public dataset considering the number of users. It contains 133 users, and 100 of them provided samples of, at least, 5 distinct sessions (Giot et al., 2009). Most sessions are spaced at 1 week at least. Each user typed the password "greyc laboratory" 12 times, on two distinct keyboards, during each session (which gives 60 samples for the 100 users having participated in each session). Both extracted features (hold time and latencies) and raw data are available (which allow computing other extracted features). The dataset[1] is stored in a sqlite dataset file.

2.2. *WEBGREYC{A,B}*

Giot et al. propose the most important public dataset (Giot et al., 2012) in terms of number of sessions (Teh et al., 2013). One hundred eighteen users had the possibility to acquire their

[1] http://www.epaymentbiometrics.ensicaen.fr/index.php?option=com_content&view=article&id=19&catid=2&Itemid=101.