Soft Biometrics Database: A Benchmark For Keystroke Dynamics Biometric Systems

Syed Zulkarnain Syed Idrus, Estelle Cherrier, Christophe Rosenberger, Patrick Bours

Soft Biometrics Database: A Benchmark For Keystroke Dynamics Biometric Systems

Syed Zulkarnain Syed Idrus[1,2], Estelle Cherrier[2], Christophe Rosenberger[2], Patrick Bours[3]

[1]Universiti Malaysia Perlis, 01000 Kangar, Perlis, Malaysia
[2]Université de Caen Basse-Normandie, UMR 6072 GREYC, F-14032 Caen, France
ENSICAEN, UMR 6072 GREYC, F-14032 Caen, France
CNRS, UMR 6072 GREYC, F-14032 Caen, France
{syed-zulkarnain.syed-idrus,estelle.cherrier,christophe.rosenberger}@ensicaen.fr
[3]NISlab, Gjøvik University College, Gjøvik, Norway
patrick.bours@hig.no

Abstract: Among all the existing biometric modalities, authentication systems based on keystroke dynamics are particularly interesting for usability reasons. Many researchers proposed in the last decades some algorithms to increase the efficiency of this biometric modality. Propose in this paper: a benchmark testing suite composed of a database containing multiple data (keystroke dynamics templates, soft biometric traits ...), which will be made available for the research community and a software that is already available for the scientific community for the evaluation of keystroke dynamics based systems. We also built the proposed biometric database on soft biometric traits for keystroke dynamics to suit the experiment. 110 people had voluntarily participated and gave their soft biometric data i.e. the way of typing, gender, age and handedness.

1 Introduction

Soft biometrics traits are physical, behavioural or biological human characteristics, classifiable in pre-defined human compliant categories, which have been derived from the way human beings normally distinguish their peers (e.g. height, gender, hair color etc ...). Those attributes have a low discriminating power, thus not capable of identification performance. Additionally, they are fully available to everyone which makes them privacy-safe. Keystroke dynamics is a viable and practical way as an addition to security for identity verification. It can be combined with passphrases authentication resulting in a more secure verification system. Soft biometrics allows a refinement of the search of the genuine user in the database, resulting in a computing time reduction. For example, if the capture corresponds to a male according to a soft biometric module, then the standard biometric authentication system can restrict its research area to male users, without considering female ones. Since this work of Jain *et al.* [JDN04], there have been several other articles dedicated to soft biometrics can be found in the literature, some of which, can be

mentioned here. The paper [AVL+06], focusing on body weight and fat measurements to enhance a fingerprint based biometric system. An overview can be found in the paper [DVDD10] about soft biometrics, under the form of a "*Bag of Soft Biometrics*": the authors make a comparison with the pioneering work of Alphonse Bertillon, whose anthropometric criteria gave rise to soft biometrics, see [Rho56]. This paper proposes some facial soft biometrics and also body soft biometrics, namely weight and clothes color detection.

Keystroke dynamics is an interesting and a low cost biometric modality as it enables the biometric system to authenticate or identify an individual based on a person's way of typing a password or a passphrase on a keyboard [GEAR09]. An original approach is presented in the work of Epp *et al.* [ELM11], strongly linked with the behavioral feature of keystroke dynamics. The authors show that it is possible to detect the emotional state of an individual through a person's way of typing. In this case, detecting anger and excitation is possible in 84% of the cases. Gender recognition is dealt in the work of Giot *et al.* in [GR12]: they show that it is possible to detect the gender of an individual through the typing of a fixed text. The gender recognition rate is more than 90% and the use of this information in association to the keystroke dynamics authentication reduces the Equal Error Rate (EER) by 20%. We also did an experiment in [SICRB13] and obtained some interesting and promising results. Our results show that given at least 10 keystroke dynamics templates of users, it is possible to detect their way of typing (using one/two hand(s)), gender, age category and handedness between 65% to 96% correct recognition accuracy performed on a dataset with five passphrases.

The objective of this paper is to present a new data collection of 110 users, both from France and Norway. This new benchmark will be released to the scientific community. We are interested in the criteria that can influence the way of typing of the users. We test if it is possible to predict if the user:

1. types with one or two hands
2. is a male or a female
3. belongs to a particular age category
4. is right- or left-handed

The aim of this paper is threefold. Section 2 deals with the state-of-the-art and the existing keystroke databases. In Section 3, it is devoted to the description of the creation of the database and its main features. The results of the analysis from the database created are discuss in Section 4.

2 State-of-the-art: Public Keystroke Dynamics Dataset

In most studies, researchers use their own dataset which, most of the time, suffers from lack of number of users and sessions. Some keystroke dynamics databases are publicly available in the literature [GEAR09, KM09, GEAR12]. In [GEAR09], several users typed the passphrase "greyc laboratory" on two different keyboards on the same computer during several sessions. 100 users have provided at least 60 samples each on 5 different sessions

spaced of one week (most of the time). In [KM09], several users have typed the password ".tie5Roanl" on a single computer during several sessions. 51 users have provided 400 samples each on 8 different sessions spaced of, at least, one day. This database contains a huge number of samples, but the time interval may be too small to track variability on a long period. These two databases are the only ones containing enough samples and users to give statistically significant results. In [GEAR12], each user was asked to key-in different logins and passwords. This is the most realistic scenario for keystroke dynamics as real users use different logins and passwords. 83 users have provided 5185 genuine samples (pair of login, password typed by its owner); 5754 impostor samples (pair of login, password typed by a user different of its owner); and 5439 imposed samples (pair of imposed login and password). This database is not the largest in terms of number of users involved, however, it is the only public keystroke dynamics providing different logins and passwords per users. Table 1 summarises this information.

Table 1: Summary of the keystroke dynamics database.

	Size	
Study	**# users**	**# samples**
[GEAR09]	100	60,000
[KM09]	51	20,400
[GEAR12]	83	5185 + 5754 / 5439
Proposed	110	11,000

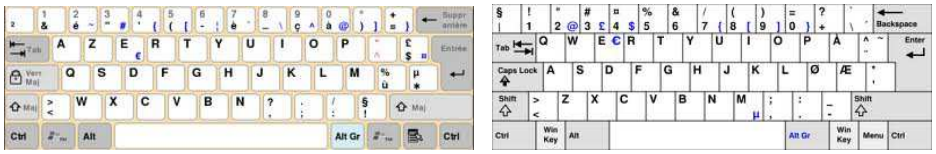
3 Benchmark Features: Biometric Database

3.1 Requirements

Hardware devices was pre-prepared such as a laptop with two external keyboards (French keyboard for users in France and Norwegian keyboard for users in Norway) i.e. AZERTY and QWERTY, respectively - the layouts are shown in Figures 1(a) and 1(b). An application to collect the keystroke dynamics data was also available. The location and position of the hardware are in a fixed position and immoveable throughout the session for the authenticity of the outcomes.

3.2 Acquisition Protocol

An experiment has been performed in two locations: France and Norway, but in fact the subjects originate from 24 different countries who are either studying or residing in one of the concerned countries. A total of 110 individuals had volunteered to participate in this



(a) French keyboard (b) Norwegian keyboard

Figure 1: Keyboards layouts

experiment where 70 of them were located in France and 40 in Norway. They are among students, researchers, faculty members, administration staff, and others. Table 2 shows the statistics repartition of gender, age and handedness among males/females and the number of each category with respect to the studied categories.

According to experts, the best password is a sentence [Dur11]. Having said that, since our study takes place both in France and Norway, we have chosen passphrases instead of sentences, well-known in both countries. Hence, for the purpose of our study for keystroke dynamics in [SICRB13], we present 5 passphrases as shown in Table 3, which are between 17 and 24 characters (including spaces) long, chosen from some of the well-known or popular names, denoted P_1 to P_5. All the participants are asked to type these 5 different passphrases 20 times. Thanks to GREYC Keystroke software developed at GREYC Laboratory (downloadable at the following address: `http://www.ecole.ensicaen.fr/~rosenber/keystroke.html`), we are able to capture biometric data. In [GEAR09], the authors describe the *GREYC-Keystroke*, which is a software developed for allowing the creation of a keystroke dynamics database and highlighted its functionalities. The keystroke application allows to add users to the application; capture the keystroke dynamics of one user several times; change the attended password; and verify the user authentication (when he/she has at least 5 captures in order to define his/her reference). Here, we define two classes of the way of typing; gender category; age category; and handedness category denoted C_1 and C_2 , respectively as follows:

- *Way of typing:* C_1 = One Hand: only one hand is used (right/left depends if the user is right/left-handed person); C_2 = Two Hands: both hands are used.
- *Gender:* C_1 = Male; C_2 = Female.
- *Age:* C_1 = $<$ 30 years old; C_2 = $\geq$ 30 years old.
- *Handedness:* C_1 = Right-handed; C_2 = Left-handed.

3.3 Keystroke Data Capture

For any keystroke capture, the captured data are the (i) code of the key, (ii) the type of event (press or release), and (iii) the time of the event. All this information is stored in

Table 2: Repartition of samples

User	70 (France); 40 (Norway)
Gender	78 males (47 from France, 31 from Norway); 32 females (23 from France, 9 from Norway)
Age Category **(between 15 and 65 years old)**	$<$ 30 years old (37 men, 14 women); $\geq$ 30 years old (41 men, 18 women)
Handedness	98 right-handed (70 men, 28 women); 12 left-handed (8 men, 4 women)

Table 3: Passphrases

Password	**Description**	**Size**
P_1	leonardo dicaprio	17-char
P_2	the rolling stones	18-char
P_3	michael schumacher	18-char
P_4	red hot chilli peppers	22-char
P_5	united states of america	24-char

the *keystroke_datas* table in the fields *rawPress* and *rawRelease*, for respectively press and release events, for each keystroke typing of an entire and correctly typed password. The data are saved following this scheme: code of the key, followed by a space, followed by the times-tamp of the event, followed by a new line and so on, for each events. The interest of storing these raw data, is to permit other researchers to create their own feature extracted data if our data does not fit their requirements. The extracted data features stored in the database are the timing differences between two events of these kinds: press/press, release/release, press/release and release/press, an additional vector resulting of the concatenation of the previous ones and the total typing timing of the password. They are stored in the fields *ppTime*, *rrTime*, *prTime*, *rpTime* and *vector* of the table *keystroke_datas* and *time_to_type* of the table *keystroke_typing*. The following are keystroke dynamics data consist of information containing the timing values of keystrokes [GEAR09], (see Figure 2):

- *ppTime (PP)*: the latencies of when the two buttons (keys) are pressed;
- *rrTime (RR)*: the latencies of when the two buttons (keys) are released;
- *prTime (PR)*: the durations of when one button (key) is pressed and the other is released;
- *rpTime (RP)*: the latencies of when one button (key) is released and the other is pressed;
- *vector (V)*: the concatenation of the four previous timing values.

The keystroke template *V* was used here for the analysis, which is the concatenation of the four mentioned timing values to perform the data analysis by classifying two classes for each category. Hence, five different features/patterns or timing vectors are extracted from each typing sample i.e. *PP, RR, PR, RP, V*. Since these extracted features are already available in the database, we can re-use them directly without having to compute it again.

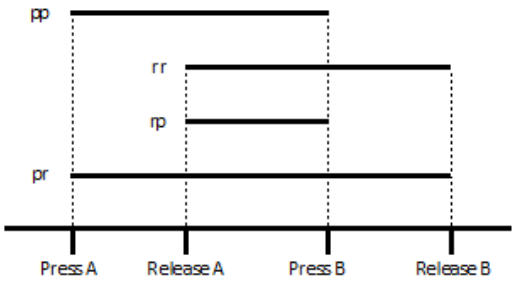


Figure 2: Keystroke typing features

3.4 Data Collection Process

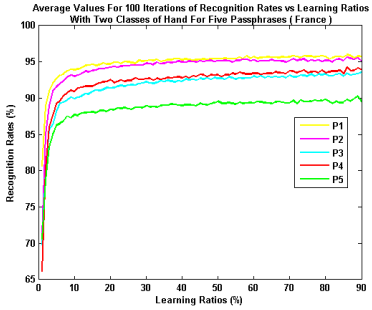
To begin the process of data acquisition, firstly, meta-information such as gender, age, handedness, and country of origin were collected. Note that also information on the keyboard is available (AZERTY or QWERTY), although that was not explicitly asked. Then, after all those information have been obtained, each user has to type each passphrase P_j, $j = 1..5$ for each hand class C_i (i = the way of typing: one/two hands), $i = 1, 2$ (1 = one hand, 2 = two hands), 10 times without errors. If there are typing errors, the current entry has to be cancelled and the user have to resume until 10 successful entries for both classes of hand have been recorded into the system. If the user is a right-handed person, he/she only need to use the right hand to key-in the passphrases in a normal typing pace, and similarly for the left-handed people. At the end of the data collection, a total of 11000 data samples (= 5 passphrases x 2 classes of hand x 110 users x 10 entries) are in the proposed biometric database. For each user, 7 out of 10 samples are used for both training and test data. The first three entries for each user are not taken into account because leeway was given to the users to allow them to train themselves for each of the given passphrases.

4 Results

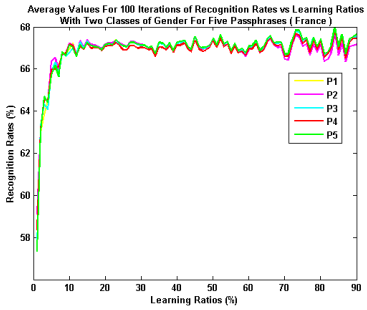
Several simulations have been performed with SVM (Support Vector Machine) for computations on several different aspects of the data namely hand recognition, gender recog-

nition, age category recognition, and handedness recognition, where the results have been published in [SICRB13]. However, we further analyse the two countries separately, both users in France and Norway, to see if there are any differences in term of their performances. Here, with substantial amount of data, we only analysed two soft biometrics information namely Hand Recognition and Gender Recognition.

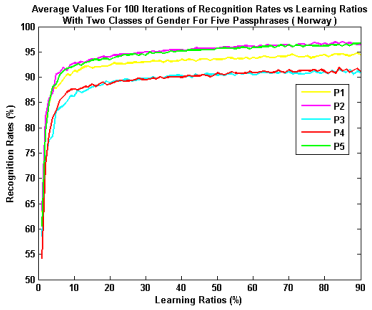
Figure 3(a) and Figure 3(c) illustrate the results of the recognition rates for France and Norway, respectively on different learning ratios with one hand (C_1) and two hands (C_2) for five different passphrases P_1 to P_5. In this experiments, the results are promising, since from the ratio of 50% of total data used for training the SVM, the recognition rate for France is between 89% and 96%, and over 90% for Norway. Figure 3(b) and Figure 3(d) illustrate the results of the recognition rates for France and Norway, respectively on different learning ratios with males (C_1) and females (C_2) for passphrases P_1 to P_5. The recognition rate, depending on the considered passphrase, is between 66.4% and 68% for France, and between 76.5% and 78.2% for Norway for a ratio over or equal to 50%.



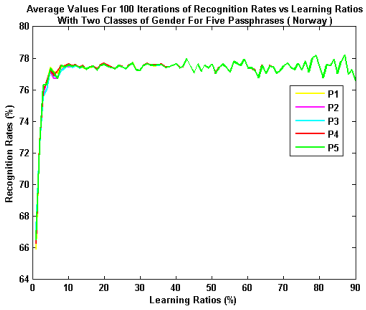
(a) France: One/two hands



(b) France: Gender



(c) Norway: One/two hands



(d) Norway: Gender

Figure 3: Average recognition rates

From the previous results, notice that the performances differ between the two soft categories because of the different criteria involved in the analysis mentioned earlier in the article. Generally, the recognition performances for all soft categories have the same trend: at the initial learning ratio, the recognition rates are quite low but then gradually increase after the learning ratios become greater i.e having more data for the learning.

5 Conclusions and Perspectives

Presented here is a new dataset for keystroke dynamics, which will soon be publicly available. This dataset is composed of several soft biometrics information of users. It consists of information on the user's way of typing by defining the number of hands used to type (one or two), gender, age and handedness. This work is however, the creation of a substantial database, with 110 users, from France and Norway, with 100 samples per user. This information could be useful and used as a reference model to assist the biometric system to better recognise a user by a way he/she types on a keyboard.

References

[AVL+06] H. Ailisto, E. Vildjiounaite, M. Lindholm, S.-M. Mäkelä, and J. Peltola. Soft biometrics–combining body weight and fat measurements with fingerprint biometrics. *Pattern Recognition Letters*, 27(5):325 – 334, 2006.

[Dur11] A. Durgahee. The best password is a sentence: says expert, May 6 2011.

[DVDD10] A. Dantcheva, C. Velardo, A. D'angelo, and J.-L. Dugelay. Bag of soft biometrics for person identification : New trends and challenges. *Mutimedia Tools and Applications, Springer, October 2010*, 10 2010.

[ELM11] C. Epp, M. Lippold, and R.L. Mandryk. Identifying emotional states using keystroke dynamics. In *Proceedings of the 2011 annual conference on human factors in computing systems*, pages 715–724, 2011.

[GEAR09] R. Giot, M. El-Abed, and C. Rosenberger. GREYC Keystroke: a Benchmark for Keystroke Dynamics Biometric Systems. *IEEE Computer Society*, 2009.

[GEAR12] R. Giot, M. El-Abed, and C. Rosenberger. Web-Based Benchmark for Keystroke Dynamics Biometric Systems: A Statistical Analysis. In *Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), 2012 Eighth International Conference on*, pages 11–15. IEEE, 2012.

[GR12] R. Giot and C. Rosenberger. A New Soft Biometric Approach For Keystroke Dynamics Based On Gender Recognition. *Int. J. Info. Tech. and Manag., Special Issue on "Advances and Trends in Biometrics by Dr Lidong Wang*, 11(1/2):35–49, 2012.

[JDN04] A.K. Jain, S.C. Dass, and K. Nandakumar. Soft biometric traits for personal recognition systems. In *Proceedings of International Conference on Biometric Authentication*, 2004.

[KM09] K. S. Killourhy and R. A. Maxion. Comparing anomaly-detection algorithms for keystroke dynamics. In *Dependable Systems & Networks, 2009. DSN'09. IEEE/IFIP International Conference on*, pages 125–134. IEEE, 2009.

[Rho56] H.T.F. Rhodes. Alphonse Bertillon: Father of scientific detection. *Pattern Recognition Letters*, 1956.

[SICRB13] S. Z. Syed Idrus, E. Cherrier, C. Rosenberger, and P. Bours. Soft Biometrics For Keystroke Dynamics. In *International Conference on Image Analysis and Recognition (ICIAR)*, June 26-28 2013.