Soft Biometrics For Keystroke Dynamics

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

Soft Biometrics For Keystroke Dynamics

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

[1]Université de Caen Basse-Normandie, UMR 6072 GREYC, F-14032 Caen, France
ENSICAEN, UMR 6072 GREYC, F-14050 Caen, France
CNRS, UMR 6072 GREYC, F-14032 Caen, France
{syed-zulkarnain.syed-idrus,estelle.cherrier,christophe.rosenberger}@ensicaen.fr
[2]NISlab, Gjøvik University College, Gjøvik, Norway
patrick.bours@hig.no

Abstract. 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. This paper presents a new soft biometric approach for keystroke dynamics. Soft biometrics traits are physical, behavioral or adhered human characteristics, 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. Thus, in this study, it consists of extracting information from the keystroke dynamics templates with the ability to recognise the hand(s) used (i.e. one/two hand(s)); the gender; the age category; and the handedness of a user when he/she types a given password or passphrase on a keyboard. Experiments were conducted on a keystroke dynamics database of 110 users and our experimental results show that the proposed methods are promising.

Keywords: Biometrics, keystroke dynamics, soft biometrics, pattern recognition, computer security.

1 Introduction

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 [5, 6]. It belongs to the class of behavioral biometrics, in the sense that the template of a user reflects an aspect of his/her behavior. Among the behavioral biometric modalities, we can mention signature analysis, gait recognition, voice recognition, or keystroke dynamics. Generally speaking, the global performances of keystroke dynamics based authentication systems are lower than the popular morphologic modalities based authentication systems (such as fingerprints, iris, *etc.*...) [11]. Besides, the main advantage of resorting to keystroke dynamics [1, 6] to authenticate a user relies in its low cost. Indeed, for this modality, no extra sensor is

required. The fact that the performances of keystroke dynamics are lower than other standard biometric modalities can be explained by the variability of the users behavior. One solution to cope with this variability is to study soft biometrics, first introduced by Jain *et al.* in [10]. In this paper "*soft biometric traits*" are defined as "*characteristics that provide some information about the individual, but lack the distinctiveness and permanence to sufficiently differentiate any two individuals*". For example Jain *et al.* consider gender, ethnicity, and height as complementary data for a usual fingerprint based biometric 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.

Concerning keystroke dynamics, an original approach is presented in the work of Epp *et al.* [4], 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 [7]: 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%. The work of Syed-Idrus *et al.* [13] show that it is possible to detect users' way of typing by using one finger (i.e. one hand) and more than one fingers (i.e. two hands) with 80% correct recognition accuracy performed on a dataset with three passwords.

The objective of this paper is twofold. First, we present a new data collection of 110 users, both from France and Norway. This new benchmark will be released to the scientific community. Then, we propose an extended study of soft biometrics for keystroke dynamics on this new database. 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

Indeed, predicting these soft features may help an authentication system (which is not considered here) to reduce the computing burden while in search of the genuine user in the database. This paper is organised as follows. Section 2 is devoted to the description of the proposed methodology: the characteristics of the database are described, together with the data collection process, and the tools that are used for analysis purposes. In Section 3, we present the obtained results while Section 4 presents the conclusions and the future works to be addressed.

2 Proposed Methodology

2.1 Biometric Database

For the purpose of this study, we have created a new biometric database. We have performed data collection and also an experiment 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 countries concerned. The selection of subjects from a myriad of nationalities would eliminate the possibility of bias sampling and thereby enhance the credibility of our experiment. Thus, a total of 110 people had volunteered to participate in this experiment where 70 of them were located in France and 40 in Norway. Table 1 shows the statistics repartition of handedness among men/women and the number of each category with respect to the other categories.

Table 1. 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)

For the creation of the biometric database, some experimentation tools are required such as a laptop; two external keyboards (French keyboard for users in France and Norwegian keyboard for users in Norway) i.e. AZERTY and QWERTY, respectively; and an application to collect the keystroke dynamics data. The location and position of the hardware are in a fixed position and immoveable throughout the session for the authenticity of the outcomes. According to experts, the best password is a sentence [3]. Hence, for the purpose of this study for keystroke dynamics, we present 5 passphrases as shown in Table 2, which are between 17 and 24 characters (including spaces) long, chosen from some of the well-known or popular names or artists (known both in France and Norway), denoted P_1 to P_5. We asked all of the participants to type these 5 different passphrases 20 times. We use the GREYC Keystroke software [5] to capture biometric data. A screenshot of this software is shown in Figure 1. 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.

Table 2. 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

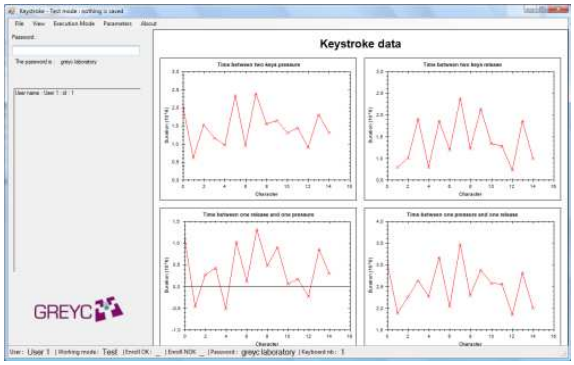


Fig. 1. GREYC keystroke software

The data that we had obtained from the 5 passphrases as listed in Table 2 are keystroke dynamics data, as mentioned earlier in the article. Keystroke dynamics data consist of information containing a field of four timing values namely: the timing pressure of when the two buttons are pressed ($ppTime$); the timing release of when the two buttons are released ($rrTime$); the timing of when one button is released and the other is pressed ($rpTime$) that is the latencies between keystrokes; and the timing of when one button is pressed and the other is released ($prTime$) that is the time durations of keystrokes [5, 6]. These data however, do not match exactly to the durations and latencies of keys because their ordering is based on time and not key code [5]. We use the keystroke template *vector*, which is the concatenation of the four mentioned timing values to perform our data analysis by classifying two classes for each category.

2.2 Data Collection Process

Each user has to type each passphrase P_j, $j = 1..5$ for each hand class C_i, $i = 1, 2$, 10 times without errors. If there are typing errors, the current entry has to be cancelled and the user will 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 will 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, we have a total of 11000 data samples (= 5 passphrases x 2 classes of hand x 110 users x 10 entries) 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 will not be taken into account because a leeway was given to the users to allow them to train themselves for each of the given passphrase.

2.3 Data Analysis

In this subsection, we present our methodology. We used a Support Vector Machine (SVM) [8, 12, 14] to perform the classification. This classifier is aimed at maximising the margins between the considered classes. From the data obtained, it takes a set of input data and predicts, for each given input, which possible class it belongs to. It consists of an evaluation of the class (C_i) recognition rate in function of the ratio of data kept for the learning stage and hence we use a Library for Support Vector Machines (LIBSVM) developed by [2] with default values. Two steps are involved: the first is devoted to a learning process, while the second is a test of the resulting SVM. The input data for the SVM that will be used to predict the 'test' data are a fixed percentage of the whole data. The remaining data are then used for test purpose, leading to a recognition rate. For data analysis, we recall that we are interested in soft biometrics criterion that can be applied to our biometric database: one or two hand(s); male or female, age < 30 or ≥ 30 years old, right/left-handed. In the sequel, we focus on the first experiment, which consists in recognising if the user types with one or two hands. The other categories are dealt with in the same manner.

The computation of SVM process is done for 100 iterations for each percentage of the learning ratio (we selected between 1% and 90% of total data to define the training set of data) and calculating the average to produce the recognition rate, which is supposed to be more consistent. Among the existing kernels for the SVM, we classically use the Radial Basis Function (RBF) kernel function. This kernel nonlinearly maps samples into a higher dimensional space so it can handle nonlinear relation between class labels and attributes. We have to search which is the best couple (C, γ) for each type of pattern. For best accuracy performance, we have to set two parameters ($C = 128.0$ is the penalisation coefficient of the SVM, $\gamma = 0.125$ is the parameter of the kernel), introduced by [9]. Here, each data file is normalised in order to have input values in the range ({-1; 1}).

To validate our performance metrics, we perform the computation of Confidence Interval (CI). A CI is necessary when it is associated with the recognition

rate of the soft biometric trait. Therefore, a CI computation may reinforce the confidence in the obtained results. It is based on a re-sampling, which consists of a random draw with a replacement of new values of example from the test base. For each draw, the data will be randomly selected i.e. different data obtain in each selection. This will be done i times in order to calculate the CI (ideally, i=1000 draws). The CI can be determined based on the percentiles of the normal distribution. It represents a measure of confidence on the estimated error rate i.e. the smaller the interval is, the more reliable the calculated error rate is. Here, the CI at 95% is defined by Equation (1), where the recognition rate is estimated from the initial sample, i the number of draws, and σ the variance of the i recognition rates calculated for the different draws.

$$CI = 1.96 \times \Sigma(rate) \pm \frac{\sigma(rate)}{\sqrt{i}} \tag{1}$$

3 Experimental Results

We had performed several simulations with SVM for computations on three different aspects of the data. The first results deal with the averaged (over 100 iterations) recognition rates for the four soft categories for different percentage of training data, from 1% to 90%. Then these results are completed by confidence intervals computation, based on a re-sampling and shuffling of the data.

• Hand Recognition

Figure 2(a) illustrates the results of the recognition rates 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 experiment, the results are promising, since from the ratio of 50% of total data used for training the SVM, the recognition rate is over 90%.

• Gender Recognition

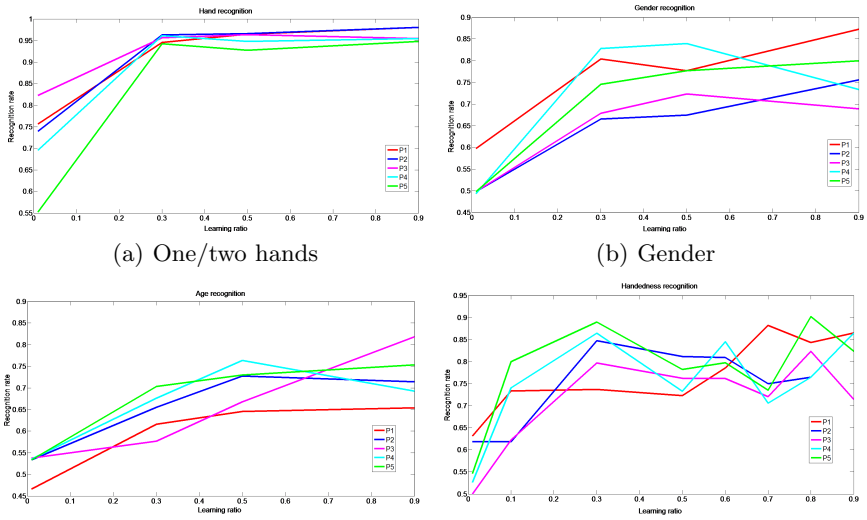
Figure 2(b) illustrates the results of the recognition rates on different learning ratios with males (C_1) and females (C_2) for passphrases P_1 to P_5. The recognition rate, depending on the considered passphrase, is between 65% and 90% for a ratio over or equal to 50%.

• Age Category Recognition

Figure 2(c) illustrates the results of the recognition rates on different learning ratios with $<$ 30 years old (C_1) and $\geq$ 30 years old (C_2) for passphrases P_1 to P_5. The recognition rate for a ratio over 50% is slightly less than that of other soft criteria, namely between 65% and 82%.

• Handedness Recognition

Figure 2(d) illustrates the results of the recognition rates on different learning ratios with right-handed (C_1) and left-handed (C_2) for passphrases P_1 to P_5. The obtained recognition rate tends to vary more than for other soft categories, but stays between 70% and 90%, which are nevertheless quite good results.



(a) One/two hands (b) Gender

(c) Age (d) Handedness

Fig. 2. Average recognition rates

• **Confidence Intervals and Discussion**

Table 3 shows the CI computed for the different categories (i.e. hand(s), gender, age, handedness): the previous results are coherent with the obtained CI. Notice that we add an extra *Gender* category: the first line corresponds to the previous simulation, with 78 males and 32 females. The second line in this category corresponds to an equilibrated case, where we have considered only 32 males to have the same number of males and females. Not surprisingly, the results are significantly better with the equilibrated classes, with an increase of 5 to 10%. We also add an extra *Age* category: < 32 and ≥ 32 years old. The results are slightly the same as those of initial age category.

Table 3. Confidence interval computation at 50% learning ratio for 5 passphrases

	P_1	P_2	P_3	P_4	P_5
Hand(s)	96% ± 0.1%	96% ± 0.1%	95% ± 0.1%	94% ± 0.1%	94% ± 0.1%
Gender (78m/32f)	64% ± 0.3%	64% ± 0.3%	63% ± 0.3%	71% ± 0.3%	68% ± 0.3%
Gender (32m/32f)	74% ± 0.3%	69% ± 0.3%	70% ± 0.2%	78% ± 0.2%	76% ± 0.2%
Age ($< 30 \geq$)	64% ± 0.2%	64% ± 0.2%	63% ± 0.2%	69% ± 0.2%	69% ± 0.2%
Age ($< 32 \geq$)	63% ± 0.2%	63% ± 0.2%	64% ± 0.2%	67% ± 0.2%	69% ± 0.2%
Handedness (12rh/12lh)	72% ± 1.2%	73% ± 1.2%	72% ± 1.2%	72% ± 1.3%	73% ± 1.2%

From the previous results, we are able to see that the performances differ from one soft category to another because of the different criteria involved in the analysis mentioned earlier in the article.

4 Conclusions and Perspectives

In this paper, we propose a new soft biometric approach for keystroke dynamics. It consists of predicting the user's way of typing by defining the number of hands used to type (one or two), the gender, the age category, and handedness, whereby the results were promising. Another part of this work is the creation of a substantial database, with 110 users, from France and Norway, with 100 samples per user. The obtained results could be used as a reference model to assist the biometric system to better recognise a user by a way he/she types on a keyboard. Hence, it would strengthen the authentication process by hindering an impostor trying to enter into the system. Having made a video recording during the data collection session, we also plan to exploit the video capture to further enhance the performances by using a fusion method as our future work. Another work in progress consists in studying the fusion of several soft categories, to enhance the recognition.

References

1. Bours, P.: Continuous keystroke dynamics: A different perspective towards biometric evaluation. Information Security Technical Report (2012) In Press, Corrected Proof.
2. Chang, C., Lin, C.: Libsvm: A library for support vector machines
3. Durgahee, A.: The best password is a sentence: says expert (May 6 2011)
4. Epp, C., Lippold, M., Mandryk, R.: Identifying emotional states using keystroke dynamics. In: Proceedings of the 2011 annual conference on human factors in computing systems. (2011) 715–724
5. Giot, R., El-Abed, M., Rosenberger, C.: Greyc keystroke: a benchmark for keystroke dynamics biometric systems. IEEE Computer Society (2009)
6. Giot, R., El-Abed, M., Rosenberger, C.: Keystroke dynamics overview. In Yang, D.J., ed.: Biometrics / Book 1. Volume 1. InTech (July 2011) 157–182
7. Giot, R., Rosenberger, C.: 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) (2012) 35–49
8. Hearst, M., Dumais, S., Osman, E., Platt, J., Scholkopf, B.: Support vector machines. Intelligent Systems and their Applications, IEEE **13**(4) (1998) 18–28
9. Hsu, C., Chang, C., Lin, C., et al.: A practical guide to support vector classification (2003)
10. Jain, A., Dass, S., Nandakumar, K.: Soft biometric traits for personal recognition systems. In: Proceedings of International Conference on Biometric Authentication,. (2004)
11. Maltoni, D., Maio, D., Jain, A., Prabhakar, S.: Handbook of fingerprint recognition. springer (2009)
12. Steinwart, I., Christmann, A.: Support vector machines. Springer (2008)
13. Syed-Idrus, S.Z., Cherrier, E., Rosenberger, C., Bours, P.: A preliminary study of a new soft biometric: finger recognition for keystroke dynamics. In: 9th Summer School for Advanced Studies on Biometrics for Secure Authentication: Understanding Man Machine Interactions in Forensics and Security Applications. (June 11-15 2012)
14. Vapnik, V.: Statistical learning theory. Wiley (1998)