Soft Biometrics for Keystroke Dynamics: Profiling Individuals While Typing Passwords

Syed Zulkarnain Syed Idrus\textsuperscript{1,2,3}, Estelle Cherrier\textsuperscript{1,2,3}, Christophe Rosenberger\textsuperscript{1,2,3}, and Patrick Bours\textsuperscript{4}

\textsuperscript{1}Université de Caen Basse-Normandie, UMR 6072 GREYC, F-14032 Caen, France
\textsuperscript{2}ENSICAEN, UMR 6072 GREYC, F-14032 Caen, France
\textsuperscript{3}CNRS, UMR 6072 GREYC, F-14032 Caen, France
\{syed-zulkarnain.syed-idrus,estelle.cherrier, christophe.rosenberger\}@ensicaen.fr
\textsuperscript{4}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 password authentication resulting in a more secure verification system. This paper presents a new profiling approach of individuals based on soft biometrics for keystroke dynamics. Soft biometrics traits are physical, behavioural or biological human characteristics, which have been derived from the way human beings normally distinguish their peers (e.g. gender, age, height, colour, race etc.). Those attributes have a low discriminating power, thus are not capable of identification performance. Additionally, they are fully available to everyone which makes them privacy-safe. In this study, it consists of extracting information from the keystroke dynamics templates with the ability to recognise the number of 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, for both known passwords and free texts. Experiments were conducted on a keystroke dynamics database of 110 users called ‘GREYC-NISLAB Keystroke’ that is available to the international scientific community. The paper proposes the application of Simple Majority Voting (SMV) and Score Fusion on the output of several Support Vector Machines (SVMs) in order to select the most suitable machine learning for keystroke dynamics based on soft biometrics features. Experimental results show that the proposed method is promising. We also present the impact of fusion schemes on the four aforementioned soft biometrics information and how fusion can enhance the overall recognition performance for both static passwords and typing rhythm of free texts i.e. digraphs.

Keywords: Biometrics, keystroke dynamics, soft biometrics, pattern recognition, data fusion, computer security.
1 Introduction

It is well-known that the way a person types on a keyboard contains timing patterns, which can be used to identify him/her and it is called keystroke dynamics. Keystroke dynamics is an interesting and a low cost biometric modality [5, 14] i.e. no extra sensor is required. 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 [13, 14]. Keystroke dynamics belongs to the class of behavioural biometrics, in the sense that the template of a user reflects an aspect of his/her behaviour. Among the behavioural biometric modalities, we can mention signature dynamics analysis, gait recognition, voice recognition, or keystroke dynamics [19, 31, 22, 30]. Generally speaking, the global performances of behavioural biometric modalities (and especially keystroke dynamics) based authentication systems are lower than the popular morphologic biometric modalities based authentication systems (such as fingerprints, iris, etc...) [25]. The fact that the performances of keystroke dynamics are lower than the other standard biometric modalities can be explained by the intraclass variability of the users behaviour. This intraclass variability pertaining to computer users is that our way of typing on a keyboard is different when we are nervous, or angry, or even sad . . . [12]. One solution to cope with this variability is to study soft biometrics, first introduced by Jain et al. in [21]. 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”. Here, Jain et al. consider gender, ethnicity, and height as complementary data for a usual fingerprint based biometric system. Thus, soft biometrics allow 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 biometrics module, then, the standard biometric identification system can confine its search area to male users, without considering female ones. Many applications on the Internet could also be addressed for marketing issues as an example.

Since the work of Jain et al., several other articles dedicated to soft biometrics can be found in the literature, some of which, will be mentioned here. An article written by Ailisto et al. [1] focuses on body weight and fat measurements to enhance a fingerprint based biometric system. An overview can be found in [8] about soft biometrics, under the form of a ‘Bag of Soft Biometrics’, where Dantcheva et al. make a comparison with the pioneering work of Alphonse Bertillon, whose anthropometric criteria gave rise to soft biometrics [35]. This paper proposes some facial soft biometrics and also body soft biometrics, namely weight and clothes colour detection. In [33], Park and Jain presents how gender or ethnicity and facial marks such as scars, moles, and freckles can be used to enhance face recognition. Dong and Woodward [11] suggested that the shape-based eyebrow features performance maybe be used for biometric recognition and soft biometric classification. In [10], Denman et al. demonstrated how an average soft biometric can be used to identify unique people in a video to calcu-
late operational measures such as the time taken to travel from point to point by considering the colour and height models for a person, where the colour model is a three part model (head, torso and legs). Marcialis et al. [26] used hair colour and ethnicity to complement face as primary biometric. In [32], Niinuma et al. proposed to use soft biometrics for continuous data authentication of a user using a computer by combining face and clothing colour with traditional face recognition and the users password. In [34], Ran et al. proposed a gait signature, consisting of several soft biometrics based on gait features, where stride length, height and gender could all be extracted from a video sequence and it was shown that these features are effective for limited recognition. According to [2], soft biometrics such as gender, age, colour, race etc. can be used to improve the speed of biometric matching through efficient filtering of the database for candidate templates, there exists no real accepted mechanisms for automatic extraction of soft biometric characteristics. Other articles related to soft biometrics focus on face recognition [20], gender recognition [23], fingerprint recognition [24] or gait recognition [31].

Concerning keystroke dynamics, an original approach is presented in the work of Epp et al. [12], strongly linked with the behavioural 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. Their most promising results are based on classifying confidence, hesitance, nervousness, relaxation, sadness, and tiredness with accuracies ranging from 77% to 88%. In this case, detecting anger and excitement is possible in 84% of the cases, where the ground truth (namely the real emotional state) is given by the user. Gender recognition is considered in the work of Giot et al. in [15]: the authors 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) of the biometric system by 20%. In [4], Bixler and D’Mello considered the possibility of automatically discriminating between natural occurrences of boredom, engagement, and neutral by analysing keystrokes, task appraisals, and stable traits of 44 individuals engaged in a writing task. They analysed by exploring several different arrangements of the data: using downsampled and/or standardised data; distinguishing between three different affect states or groups of two; and using keystroke/timing features in isolated or coupled with stable traits and/or task appraisals. Their results indicated that the use of raw data and the feature set that combined keystroke/timing features with task appraisals and stable traits, yielded accuracies that were 11% to 38% above random guessing and were generalised to new individuals. Our recent study in [40] also showed that it is possible to detect users’ way of typing by using one/two hand(s) with over 90% recognition rate; gender between 65% and 90%; age category between 65% and 82%; and handedness between 70% and 90% correct recognition accuracy with 110 users.
Several other studies on keystroke dynamics have also been conducted recently. Hence, keystroke and stylometry (an application of the study of linguistic style) behavioural biometrics were investigated by [29]. The authors are interested to verify the identity of students in online examination environments and compare the performance statistics between the two modalities. Their results showed that on the test input of 500 and 1000 words, the performance of the keystroke system was at 99.96% and 100.00%, while the stylometry system was much weaker at 74% and 78%, respectively. The authors in [28] are interested in authenticating online student test takers and computer users in security sensitive environments. The authentication process, however, operates on keystroke data windows was as short as $\frac{1}{2}$ minute. The performance on 14, 30, and 119 users was 99.6%, 98.3%, and 96.3%, respectively, on 755-keystroke samples. The keystroke biometric classification system described in [3] evaluates two types of short input: password and numeric keypad input. On the password input, the EER is equal to 8.7%, whereas on the numeric keypad input, the system achieved an EER that is equal to between 6.1% and 10.5%.

The objective of this paper is to propose an extended study of soft biometrics for keystroke dynamics from our previous study in [40] on a new biometric benchmark database called ‘GREYC-NISLAB Keystroke’ [39] that we have created. We propose in this paper a thorough evaluation of the soft biometrics system and a comparison between static passwords and free texts (digraphs). We also show how the performances can be increased significantly by data fusion. 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

This paper is organised as follows. Section 2 is devoted to the description of the proposed methodology. In Section 3, we describe the protocol that we applied and present the obtained results on the benchmark database in Section 4. Section 5 presents the conclusions and the different perspectives of this study.

2 Proposed Methodology

In general, the proposed keystroke dynamics authentication system involves a keyboard and an application for the capture and processing of the biometric information. Users are required to type on a keyboard with a dedicated software to capture the keystroke of a given password/passphrase. Each capture is stored in a database within the application in the form of keystroke/timing features for all correct and incorrect entries. These features are composed of several timing values that are extracted, which is the pattern vector that is used for the analysis. Once we have all the necessary features extracted, two steps are performed i.e. one for the learning; one for the recognition. The learning involves the training.
of an SVM [41] and the recognition amounts to the computing the accuracy rate of the prediction (i.e. learning data compared to the actual data). The prediction results are the results of the decision, whether the SVM has predicted that the unknown record (data) belongs to a particular class or another to its nearest neighbour. A graphical representation of the overall process is illustrated in Figure 1. In order to further enhance the overall recognition performance, data fusion is then applied. When applying fusion, we will show that the accuracy rate significantly improves. Therefore, the final decision relies on the soft biometric information that has better performance i.e. the higher the rate, the better the performance.

2.1 Data Capture

From a pattern recognition perspective, a pattern (in this case, the keystroke timings) is useful only to the extent that one or more features may be extracted from it to successfully discriminate between one user and another. Thus, different types of features can be extracted from a user while typing on a keyboard i.e. the timing pattern of keystrokes: (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 the keystroke data 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 in this following scheme: code of the key; followed by a space; followed by the time-stamp 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. Keystroke dynamics features are usually extracted using the timing information of the key down/hold/up events: the timing of when the two buttons are pressed (ppTime); the timing of when the two buttons are released (rrTime); the timing of when one button is pressed and the other is released (prTime) that is the time durations of keystrokes; the timing of when one button is released and the other is pressed (rpTime) that is the latencies between keystrokes; an additional vector \( V \) resulting of the concatenation of the previous ones; and the total typing timing of the password [13, 14]. The following are keystroke dynamics data consist of information containing the timing values of keystrokes (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 \( V \) is use here for the analysis, which is the concatenation of the four mentioned timing values to perform the data analysis by
Fig. 1. The overall process of the proposed system.
classifying two classes for each category. Hence, five different features/patterns or timing vectors are extracted from each typing sample \textit{i.e.} PP, RR, PR, RP, \textit{V}. Since these extracted features are already available in the database, we can reuse them directly without having to compute it again.

![Keystroke typing features](image)

Fig. 2. Keystroke typing features.

For keystroke dynamics systems, there are two approaches that we use here namely password and free text. With password, we analyse all the typing features for each known and fixed password. With free text, the analysis is based on digraphs, which are the time latencies between two successive keystrokes \textit{i.e.} digraphs transition time. These typing rhythms from the texts typed by a user without any specific constraint is taken into account.

2.2 Data Analysis

In this subsection, we present the data analysis methodology. We use an SVM \cite{17, 38, 41} to perform the classification. This classifier is aimed at maximising the margins between the considered classes \( C_i \) (see Figure 3. Hence, we use a library for SVMs (LIBSVMs) developed by \cite{6} with default values. From the obtained data, 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. There are two steps involved: the first is devoted to a learning process, while the second is a test of the resulting SVM. The portion of the input data for the SVM is taken to train the learning process at a fixed percentage of the whole data. The remaining data is then used for test purposes, leading to a correct 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 \( \geq 30 \) years old, right- or left-handed. In the sequel, we focus on the first experiment, which consists in recognising if the user types with one or two hands. Indeed, the other categories are dealt with in a similar manner.
The computation of the SVM process is done for 100 iterations for each percentage of the learning ratio (we select from 1% to 90% of the total data to define the training set) and calculating the average over the 100 iterations 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 with default values because of its good general performance and only a few number of parameters (only two: $C$ and $\gamma$). This kernel nonlinearly maps samples into a higher dimensional space so it can handle nonlinear relations between class labels and attributes i.e. when we cannot find a linear separator, data points are projected into an (usually) higher-dimensional space where the data points effectively become linearly separable (this projection is realised via kernel techniques) [27]. There are several parameter couples that can be use with this kernel, however, in order to maximise the performance, we have to determine which is the best couple for our computation i.e. $(C, \gamma)$ for each type of pattern. Therefore, for best accuracy performance, we set these two parameters ($C = 128$ is the penalisation coefficient of the SVM; $\gamma = 0.125$ is the parameter of the kernel), as introduced by [18]. Here, each data file is normalised in order to have input values in the range $\{-1; 1\}$.

2.3 Data Fusion Process

Data fusion, is the process of integration of multiple data and knowledge representing the same real-world object into a consistent, accurate, and useful representation. Data fusion processes are often categorised as low, intermediate or high, depending on the processing stage at which fusion takes place [16]. Low
level data fusion combines several sources of raw data to produce new raw data. The expectation is that fused data is more informative and synthetic than the original inputs. In a simple case, all attributes are uniform across the entire analysis domain, so attributes may be simply assigned. In more realistic applications, attributes are rarely uniform and some type of interpolation is usually required to properly assign attributes to the data points in the fused set.

In this study, for the data fusion, we apply two techniques based on ‘majority voting’ and ‘score fusion’ with binary classifications as illustrated in Figure 4. We take ‘gender’ for example, and perform similar approaches for the other soft biometrics information. We select a set of users \( \{u_1 ..., u_i\} \) from the male data, and similarly for the female where both sets of data are equal in size for all 5 passphrases. We retain the same set of users for each passphrase. Before each data selection \( \{u_1 ..., u_i\} \), the data are randomly selected (i.e. shuffled at each selection). Then, we perform an SVM with 1 iteration. Once the SVM is completed, several data are produced namely \( \text{train}, \text{test} \) and \( \text{result} \). The \( \text{train} \) contains data that are used/selected for the learning (training). Here, we use a ratio of 50% i.e. 50% are dedicated for learning and the remaining 50% of the data are used for testing. The SVM computes the prediction from the learning data with the testing data. The \( \text{result} \) from this computation produces the predicted class labels i.e. \(-1\) = female or ‘1’ = male and its probability values data i.e. \(0.16\), where 16% belongs to class ‘-1’ or ‘0.84’, where, 84% belongs to class ‘1’). So, in this case, the result of predicted class label is ‘1’, where SVM predicted at 84% that it is ‘male’ (see Figure 5).

Using the first method, we compute the ‘majority voting’ i.e. the predicted label \( \{-1; 1\} \). We add each row of the results data on all 5 passphrases all together. If the values are greater than 0, then we assign those values with 1, and for the others -1. The same process is performed on the test data. Thus, we only have the values \( \{-1; 1\} \) for both the result and test data.

For the second method, instead of using the predicted label \( \{-1; 1\} \), we compute the ‘score fusion’. We compute this score by using the guessed label and its probability by adding each row of the probability results data on all 5 passphrases all together. In this method, we obtain a score in the range \([0 \text{ to } 1]\). Gender, for example, scores closer to 0 represent a ‘female’ (i.e. -1), while scores closer to 1 represent a ‘male’ (i.e. 1), and similarly for the other soft biometrics information.

Equation 1 presents the way of computing both ‘majority voting’ and ‘score fusion’, where \( \text{predict} \) presents the ‘predicted label’ (e.g. -1 or 1) and \( \text{probability} \) is its ‘probability value’ (e.g. 0.16 or 0.84).

\[
SVM = \text{predict} \times \text{probability} 
\] (1)
To classify an unknown record (data):

(i) Compute distance to other training records;
(ii) Identify k nearest neighbours;
(iii) Use class labels of nearest neighbours to determine the class label or the probability value of unknown record e.g. by taking its ‘majority voting’ or ‘score fusion’
   - Majority voting: -1 / 1
   - Score fusion: 0.16 / 0.84

- 84% of being class 1 (+), 16% of being class -1 (-);
- here, we are sure at 84% that the unknown record belongs to class 1 i.e. ‘male’ because its nearest neighbour is ‘+’;
- but, say, if the value is 58%, we cannot be so certain which class it belongs to i.e. near the border line.

Fig. 5. Majority voting and score fusion for k nearest neighbours based on gender.
Once this process has been completed, we can compute the confusion matrix to obtain the recognition accuracy rate i.e. ground truth. In the field of machine learning, a confusion matrix is a specific table layout that allows visualisation of the performance of an algorithm, typically a supervised learning one (in unsupervised learning it is usually called a matching matrix). Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. The name stems from the fact that it makes it easy to see if the system is confusing two classes (i.e. commonly mislabeling one as another).

3 Experimental Protocol

In this paper, we briefly describe the protocol that we applied. For the interested user, the details have been published and can be obtained in [39]. Based on the results of this study, we created a new biometric benchmark database called ‘GREYC-NISLAB Keystroke’. The main interests of such a database are to avoid researchers to spend too much time for creating it, and to easily compare the performance of different algorithms with the same input data (reproducible research). In [40], an experiment has been performed in two locations: France and Norway, and a total of 110 individuals had volunteered to participate. We use two desktop keyboards (French keyboard for users in France and Norwegian keyboard for users in Norway) i.e. AZERTY and QWERTY (it is not a classical QWERTY keyboard, as we do not use specific Norwegian keys), respectively.

During the data acquisition, some metadata such as gender, age, handedness, and country of origin were collected. We have chosen passphrases that are well-known in both countries, which are between 17 and 24 characters (including spaces) long. All the participants are asked to type 5 different passphrases 20 times (10 times with one hand and 10 times with two hands), as shown in Table 1. Thanks to GREYC Keystroke software developed at GREYC Laboratory (downloadable from the following address: http://www.ecole.ensicaen.fr/~rosenber/keystroke.html), we are able to capture biometric data. 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 benchmark database. For each user, 7 out of 10 samples are used for training and testing 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. We define two classes $C_1$ and $C_2$ for each category: hand category (the way of typing); gender category; age category; and handedness category as follows:

- **Hand category**: $C_1 =$ 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 category**: $C_1 =$ Male; $C_2 =$ Female.
• **Age category:** $C_1 = < 30$ years old; $C_2 = \geq 30$ years old.

• **Handedness category:** $C_1 =$ Right-handed; $C_2 =$ Left-handed.

Here, for hand category, we use all the data. Whereas for the other soft biometrics information, we only use data corresponding to two hands *i.e.* the usual/normal way of typing on a keyboard.

### Table 1. Passphrases.

<table>
<thead>
<tr>
<th>Label</th>
<th>Description</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_1$</td>
<td>leonardo dicaprio</td>
<td>17-char</td>
</tr>
<tr>
<td>$P_2$</td>
<td>the rolling stones</td>
<td>18-char</td>
</tr>
<tr>
<td>$P_3$</td>
<td>michael schumacher</td>
<td>18-char</td>
</tr>
<tr>
<td>$P_4$</td>
<td>red hot chilli peppers</td>
<td>22-char</td>
</tr>
<tr>
<td>$P_5$</td>
<td>united states of america</td>
<td>24-char</td>
</tr>
</tbody>
</table>

To validate the proposed recognition system, 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 will 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 are randomly selected *i.e.* different data are extracted in each selection. This is done 100 times in order to calculate the CI. 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 2, where the recognition rate is estimated from the initial sample and $\sigma$ the variance of the 100 recognition rates calculated for the different draws.

$$CI = 1.96 \times \Sigma(rate) \pm \frac{\sigma(rate)}{\sqrt{100}}$$

Subsequently, we perform a distance measure to consider the different timing information between two-character sequences known as *digraphs*. Digraphs are the latency times between two successive keystrokes. We extract the keystroke features using the mean and variance of digraphs. Since some digraphs are more frequent, the size of the timing vector may differ according to respective digraph [9]. It was shown in [36] that the typing pattern of a letter sequence may change
when it is part of a larger word. For example, digraph ‘IS’ has different timing information in typing the word ‘IS’ and the word ‘FUTURISTIC’.

In a long text, there may be more than one instance of a digraph. However, the mean of all these instances is used as a corresponding latency time of those digraphs. Here, the digraphs are between one and four occurrences from the 26 letters in the English alphabet as shown in Figure 6. Thus, there are 39 with one occurrence, 11 with two occurrences, 2 with three occurrences, and 1 with four occurrences. Therefore, here, the latency times of all digraphs in a text sample are not used as the features of that sample. We are interested and analyse only the mean of the occurrences that are equal or greater than two and hence, there are a total of 14 occurrences that fall in this category namely the digraph latency of ‘ca’, ‘ic’, ‘ed’, ‘he’, ‘pe’, ‘te’, ‘ch’, ‘li’, ‘ri’, ‘ll’, ‘on’, ‘er’, ‘es’ and ‘st’.

| a | b | c | d | e | f | g | h | i | j | k | l | m | n | o | p | q | r | s | t | u | v | w | x | y | z |
| s | g | h | i | j | k | l | m | n | o | p | q | r | s | t | u | v | w | x | y | z | s | g | h | i | j |

**Fig. 6.** Digraphs and its number of occurrences.

Finally, we compute the confusion matrix in order to obtain the recognition accuracy rate *i.e.* ground truth. As mentioned earlier, the confusion matrix is often called the contingency table or the error matrix [37], where for each class, it presents the percentage of correctly classified users, so it gives more precise details than the overall EER. Below we show an example of how we compute the confusion matrix for gender, and similarly for the other soft biometrics in-
formation. We define:

<table>
<thead>
<tr>
<th></th>
<th>One Hand ($I$) = 1</th>
<th>Two Hands ($2$) = -1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male ($M$)</td>
<td>1</td>
<td>Female ($F$) = -1</td>
</tr>
<tr>
<td>&lt; 30 years old ($&lt;$ 30) = 1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>≥ 30 years old ($\geq$ 30) = -1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Right-handed ($RH$)</td>
<td>1</td>
<td>Left-handed ($LH$)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>= -1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Predicted Data</th>
<th>Real Data</th>
<th>Male ($M$): [1]</th>
<th>Female ($F$): [-1]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male ($M$): [1]</td>
<td></td>
<td>$a$ (1/1)</td>
<td>$b$ (-1/1)</td>
</tr>
<tr>
<td>Female ($F$): [-1]</td>
<td></td>
<td>$c$ (1/-1)</td>
<td>$d$ (-1/-1)</td>
</tr>
</tbody>
</table>

The basic performance measure:

\[
a = (1/1); \]
where, ‘real data’ = $M$ and ‘predicted data’ = $M$, hence correctly predicted $M$ by SVM i.e. True Positive (TP).

\[
b = (-1/1); \]
where, ‘real data’ = $F$ and ‘predicted data’ = $M$, hence wrongly predicted $F$ as $M$ by SVM i.e. False Positive (FP).

\[
c = (1/-1); \]
where, ‘real data’ = $M$ and ‘predicted data’ = $F$, hence wrongly predicted $M$ as $F$ by SVM i.e. False Negative (FN).

\[
d = (-1/-1); \]
where, ‘real data’ = $F$ and ‘predicted data’ = $F$, hence correctly predicted $F$ by SVM i.e. True Negative (TN).
We compute the total data by the following equations:

\[ \text{total\_number\_of\_male} = a + c \quad (3) \]

\[ \text{total\_number\_of\_female} = b + d \quad (4) \]

\[ \text{total\_data} = (a + c) + (b + d) \quad (5) \]

To compute the recognition rate \((r)\), we take the two values that are correctly predicted for both \(M\) and \(F\) i.e. \(a\) and \(d\), add them together and divide by the total number of data. This is defined by:

\[ r = \frac{a + d}{\text{total\_data}} \times 100\% \quad (6) \]

Equation 6 gives us the percentage of the recognition rate and basically, the higher the accuracy (recognition) rate, the better the performance is.

4 Experimental Results

In the first subsection, we quantify the performance results of soft biometrics for keystroke dynamics with known passwords \(i.e.\) when the user types a specific password). Subsequently, in the second subsection it is not too confined, with any combinations of two-key characters (digraphs) with free text, the user can type arbitrary text as input without any specific constraints \(i.e.\) user is free to type whatever he/she wants, rather than be constrained to a pre-determined text). The last subsection illustrates the results from the previously introduced techniques that can enhance the performance of soft biometrics for keystroke dynamics for both static texts (known passwords) and free text with ‘majority voting’ and ‘score fusion’.

4.1 Password: Static Texts

We performed several simulations with SVM for computations on two different aspects of the data. Firstly, the results deal with the averaged (over 100 iterations) recognition rates by removing the first three entries for the four soft categories for different percentage of training data and select from 1\% to 90\% of the ratios \(i.e.\) 7 out of 10 samples are used (as mentioned earlier). Then, these results are completed by confidence intervals computation, based on a re-sampling and shuffling of the data.
• **Hand Category Recognition**

Figure 7 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 $P_5$. The data distribution (key data) to evaluate these results is 100% of the data are used (i.e. $C_1 = 110$ users’ samples that use one hand; $C_2 = 110$ users’ samples that use two hands). 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 Category Recognition**

Figure 8 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 data distribution (key data) to evaluate these results is about 30% of the data are used in order to have equilibrated classes (i.e. $C_1 = 32$ users’ samples that are male; $C_2 = 32$ users’ samples that are female). The recognition rate, depending on the considered passphrase, is between 70% and 86% for a ratio over or equal to 50%.

• **Age Category Recognition**

Figure 9 illustrates the results of the recognition rates on different learning ratios with < 30 years old ($C_1$) and ≥ 30 years old ($C_2$) for passphrases $P_1$ to $P_5$. The data distribution (key data) to evaluate these results is about 46% of the data are used in order to have equilibrated classes (i.e. $C_1 = 51$ users’ samples that are < 30; $C_2 = 51$ users’ samples that are ≥ 30). The recognition rate for a ratio over 50% is slightly less than that of other soft criteria, namely between 67% and 78%.

• **Handedness Category Recognition**

Figure 10 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 data distribution (key data) to evaluate these results is about 10% of the data are used in order to have equilibrated classes (i.e. $C_1 = 12$ users’ samples that are right-handed; $C_2 = 12$ users’ samples that are left-handed). The obtained recognition rate tends to vary more than for other soft categories, but stays between 76% and 88%, which are nevertheless quite good results. However, as mentioned, we select only a part of the right-handed users to have the same number of left-handed and right-handed in this experiment.

• **Confidence Intervals**

Table 2 shows the CI computed at 50% learning ratio for different categories (i.e. hand, 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. As expected, the results are significantly better with the equilibrated classes, with an increase of
Fig. 7. Average values for 100 iterations of recognition rates at 1% to 90% learning ratios with two classes of hand for five passphrases.

Fig. 8. Average values for 100 iterations of recognition rates at 1% to 90% learning ratios with two classes of gender for five passphrases.
Fig. 9. Average values for 100 iterations of recognition rates at 1% to 90% learning ratios with two classes of age for five passphrases.

Fig. 10. Average values for 100 iterations of recognition rates at 10% to 90% learning ratios with two classes of handedness for five passphrases.
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 2. Confidence interval computation at 50% learning ratio for 5 passphrases and the data distribution in each class.

<table>
<thead>
<tr>
<th></th>
<th>$P_1$</th>
<th>$P_2$</th>
<th>$P_3$</th>
<th>$P_4$</th>
<th>$P_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Hand</strong></td>
<td>96% ± 0.1%</td>
<td>96% ± 0.1%</td>
<td>95% ± 0.1%</td>
<td>94% ± 0.1%</td>
<td>94% ± 0.1%</td>
</tr>
<tr>
<td>$(C_1=110; C_2=110)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td>64% ± 0.3%</td>
<td>64% ± 0.3%</td>
<td>63% ± 0.3%</td>
<td>71% ± 0.3%</td>
<td>68% ± 0.3%</td>
</tr>
<tr>
<td>$(C_1=78; C_2=32)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td>74% ± 0.3%</td>
<td>69% ± 0.3%</td>
<td>70% ± 0.2%</td>
<td>78% ± 0.2%</td>
<td>76% ± 0.2%</td>
</tr>
<tr>
<td>$(C_1=32; C_2=32)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Age</strong></td>
<td>64% ± 0.2%</td>
<td>64% ± 0.2%</td>
<td>63% ± 0.2%</td>
<td>69% ± 0.2%</td>
<td>69% ± 0.2%</td>
</tr>
<tr>
<td>$(C_1=51; C_2=51)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Age</strong></td>
<td>63% ± 0.2%</td>
<td>63% ± 0.2%</td>
<td>64% ± 0.2%</td>
<td>67% ± 0.2%</td>
<td>69% ± 0.2%</td>
</tr>
<tr>
<td>$(C_1=51; C_2=51)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Handedness</strong></td>
<td>72% ± 1.2%</td>
<td>73% ± 1.2%</td>
<td>72% ± 1.2%</td>
<td>72% ± 1.2%</td>
<td>74% ± 1.2%</td>
</tr>
<tr>
<td>$(C_1=12; C_2=12)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4.2 Free Text: Digraphs

We performed similar simulations with SVM for computations as mentioned in Section 4.1. The first results deal with the averaged (over 100 iterations) recognition rates by removing the first three entries (i.e. 7/10) for the four soft categories for different percentage of training data, from 1% to 90%.

- **Soft Biometrics Information Recognition Based on Free Text**

Figure 11 illustrates the results of the recognition rates on different learning ratios with $C_1$: one hand, male, age < 30 years old, right-handed; and $C_2$: two hands, female, age ≥ 30 years old for all four different soft biometrics information. In this experiment, the results are promising, from the ratio of 50% of total data used for training the SVM, the recognition rate for Hand Category Recognition is much over 90%; Gender Category Recognition is between 79% and 84%; Age Category Recognition is between 72% and 75%; and Handedness Category Recognition is between 83% and 88%. Table 3 summarises the performance comparison of recognition rate between password and free text from 50% to 90% learning ratio.

4.3 Confusion Matrix: Majority Voting and Score Fusion for Password and Free Text

In order to further enhance the performance, we perform the data fusion method, where we will show that there is a great increase in the accuracy (recognition)
rate results. For password, by applying Equation 6, the results of our confusion matrix have improved significantly by fusing the data on all soft biometrics information at 50% learning ratio based on ‘static texts’. We apply the same equation and ratio for free text and the results of our confusion matrix are based on ‘diagrams’. Then, we perform the performance comparison: (i) before fusion; (ii) fusion based on ‘majority voting’; and (iii) fusion based on ‘score’ for password and free text. Table 4 summarises this information.

![Recognition Rates of Two Classes of Soft Biometrics Information on Free Text (digraphs)](image)

**Fig. 11.** Average values for 100 iterations of recognition rates at 1% to 90% learning ratios with two classes of soft biometrics information with 14 digraphs (occurrences ≥ 2) on free text.

**Table 3.** Summary of performance comparison of recognition rates between password and free text at 50% to 90% learning ratios.

<table>
<thead>
<tr>
<th></th>
<th>Password</th>
<th>Free Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hand (C₁=110; C₂=110)</td>
<td>&gt; 90%</td>
<td>&gt; 90%</td>
</tr>
<tr>
<td>Gender (C₁=32; C₂=32)</td>
<td>b/t 70% and 86%</td>
<td>b/t 79% and 84%</td>
</tr>
<tr>
<td>Age ([C₁=51] &lt;30&gt; [C₂=51])</td>
<td>b/t 67% and 78%</td>
<td>b/t 72% and 75%</td>
</tr>
<tr>
<td>Handedness (C₁=12; C₂=12)</td>
<td>b/t 78% and 88%</td>
<td>b/t 83% and 88%</td>
</tr>
</tbody>
</table>
Table 4. Performance comparison before and after fusion between password and free text at 50% learning ratio.

<table>
<thead>
<tr>
<th>Technique</th>
<th>Soft Biometric Information</th>
<th>SVM 1 (Before Fusion)</th>
<th>SVM 2 (Majority Voting)</th>
<th>SVM 3 (Score Fusion)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Password</strong></td>
<td>Hand Category</td>
<td>93.66%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>Gender Category</td>
<td>62.5%</td>
<td>85.71%</td>
<td>92.14%</td>
</tr>
<tr>
<td></td>
<td>Age Category</td>
<td>55.49%</td>
<td>86.67%</td>
<td>85.71%</td>
</tr>
<tr>
<td></td>
<td>Handedness Category</td>
<td>61.65%</td>
<td>84.52%</td>
<td>91.67%</td>
</tr>
<tr>
<td><strong>Free Text</strong></td>
<td>Hand Category</td>
<td>96.57%</td>
<td>96.57%</td>
<td>96.57%</td>
</tr>
<tr>
<td></td>
<td>Gender Category</td>
<td>80%</td>
<td>80%</td>
<td>80%</td>
</tr>
<tr>
<td></td>
<td>Age Category</td>
<td>65.71%</td>
<td>65.71%</td>
<td>65.71%</td>
</tr>
<tr>
<td></td>
<td>Handedness Category</td>
<td>83.33%</td>
<td>83.33%</td>
<td>83.33%</td>
</tr>
</tbody>
</table>

4.4 Discussions

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. 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.

Therefore, by fusing using both ‘majority voting’ and ‘score fusion’ techniques have significantly increased the recognition accuracy performance rate on all soft biometrics information from the initial results (*i.e.* before fusion) between 23% to 31%. ‘Score fusion’, however, is a better way of fusing the data with password (static texts), where we can see the results of this performance had increased significantly by up to 30% of the recognition accuracy rates.

Free text, on the other hand, the recognition performance rates are somewhat less better than password on all soft biometrics information as illustrated in Table 4. Nonetheless, with any combinations of two-key characters and without any specific constraints for the input text, the results of free text performance are also quite favourable depending on the considered soft biometrics.
5 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, where the results were promising for both password and free text. We are also able to enhance the soft biometrics recognition rate significantly by data fusion and achieve higher accuracy performance. Another part of this work is a comparative study between password and free text, where both approaches give good results. For password, it is based on static texts that the users are obliged to type with some constraints i.e. users type specific pre-enrolled strings. Free text, on the other hand, with any combinations of two-key characters (digraphs) is also able to provide good recognition rates without any specific constraints to the users when typing i.e. users can type arbitrary text as input, in which the user to be authenticated is free to type whatever he/she wants, rather than be constrained to a pre-determined text [36].

Thus, in our study on soft biometric approach for keystroke dynamics, we found that password is good for recognition if we want to impose, say, a challenge for the user to type. Free text, on the other hand, is best for recognition if a set of texts or passwords are created by the user him/herself, however, the effectiveness of digraphs are discriminative only when they are word-specific i.e. digraph features depend on the word context they are computed in [42]. Therefore, 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, not only it would strengthen the authentication process by hindering an impostor trying to enter into the system, but also cut down on the computation time.

The perspectives of this work concern the use of all these soft biometrics information in order to improve the recognition of users, where the verification score could combine two information. The first one corresponds to the score provided a biometric system when comparing the reference and a capture template. The second one could be a reliability index by verifying the concordance between one extracted soft biometric information (such as gender) and the known information. However, many other applications of this work concern social networks monitoring in the sense that it could be possible to analyse when user is chatting if the extracted soft biometrics information corresponds to its known profile.

References

28. Monaco, J.V., Bakelman, N., Cha, S.H., Tappert, C.C.: Recent advances in the development of a long-text-input keystroke biometric authentication system for arbitrary text input. studies 17 (2013) 25
29. Monaco, J.V., Stewart, J.C., Cha, S.H., Tappert, C.C.: Behavioral biometric verification of student identity in online course assessment and authentication of authors in literary works