Can Chronological Information Be Used As A Soft Biometric In Keystroke Dynamics ?

Romain Giot and Christophe Rosenberger
Université de Caen, UMR 6072 GREYC
ENSICAEN, UMR 6072 GREYC
CNRS, UMR 6072 GREYC
Email: romain.giot@ensicaen.fr
Email: christophe.rosenberger@ensicaen.fr

Bernadette Dorizzi
Institut Télécom; Télécom SudParis
UMR 5157 SAMOVAR
Email: bernadette.dorizzi@it-sudparis.eu

***Abstract*—Keystroke dynamics is a behavioral biometric modality which uses typing patterns on a keyboard to recognize individuals. The way of typing the password slightly changes with time, because of various factors (including user's training). This modification in the way of typing results in a decrease of performance recognition over time. In this paper, we analyse the correlation between the comparison score between a query, and a reference and the number of times the user has typed the password. After having quantified this correlation, we analyse the possibility of using stacked classification to take this aspect in consideration during authentication. Then, we verify if it is possible to classify users on their way of evolving their typing pattern. Results show that even if comparison scores are correlated with the number of time the user has typed the password, the use of a stacked classifier does not improve the results much.**

I. Introduction

It is known that performance of biometric systems degrades with time. This can be explained by various facts: acquisition problems for enrolled samples (*i.e.* model not representative of the user), template ageing, lack of samples for enrollment (*i.e.*, impossibility to accurately model the user), or, for behavioral biometrics, training on the action to do when authenticating. These variations may be permanent or temporary. Several corrective actions can be used to overcome this problem: capture of several samples during enrollment, multiple tries during authentication [1], selection of features which do not vary a lot over time [7], artificially ageing of the template for matching the age of the query [11], evolution of the decision frontier [3] or updating of the biometric reference [10]. Most papers do not respect the chronology of the captured samples when evaluating their biometric reference update methods. It may not be too problematic for morphological biometrics, but, we think it is not realistic for behavioral biometrics where a personal aspect must be taken into account: the training of the user while performing the authentication gesture.

The aim of this paper is to verify if we must respect the chronology for keystroke dynamics, and, try to use this information to evolve the decision frontier of acceptance. Our contributions are (i) the statistical analysis of score variation over time for keystroke dynamics authentication, and (ii) the proposition of a multimodal stacked classifier for keystroke dynamics authentication, and its evaluation. We also want to analyze users' recognition performance depending on their way of varying the way of typing over time. Section II presents the experimental protocol. Section III presents the results of the various studies. Section IV concludes this paper.

II. Experimental protocols

In this section, we present the common parameters (databases and keystroke dynamics authentication methods) of each scenario, the experimental protocols we settled for analysing the score evolution over time, and the definition of the multimodal stacked classifier.

A. Experimental Protocol

We have used two different keystroke dynamics datasets: DSN2009 [9] (51 users, 400 samples per user, 8 sessions, password: ".tie5Roanl") and GREYC [5] (100 users, 60 samples per user, 5 sessions, password: "greyc laboratory"). We have used two different statistical authentication methods [2], [6], and one based on rhythm (each dimension of the samples is discretized in a 5 characters alphabet according to its deviation to the mean value related to the standard deviation, the score is the Hamming distance between the query and the mean of enrolled samples after discretisation). We have also implemented a weighted sum score fusion function after normalizing scores with the *z-score* technique. For each database, we use the first session for training the model. The other sessions serve for the validation process.

B. Score Evolution Over Time

It has already been shown that using template update mechanisms improves the recognition performance of keystroke dynamics systems, both in supervised scenario [8], and in semi-supervised scenario [4]. We have also shown that there is a stabilisation in the typing pattern with time [4] (the more we type, the more similar the samples are). However, the evolution of scores while template is ageing has never been studied for keystroke dynamics (although we can find such kind of study for face recognition [3]). We think that template update helps incorporating missing intraclass variations for morphological modalities (fingerprint, face, ...), while it helps taking into

account the user learning of the action for behavioral modalities (keystroke dynamics, signature recognition). Thus, we think that there is a link between the recognition score and the number of times the user interacts with the system (in our case, the password typing). That is why we have annotated each keystroke sample by its creation order from its owner (*i.e.*, each sample $\mathbf{s}$ is ordered: $\mathbf{s}_i^j$ belongs to user i and has been captured before $\mathbf{s}_i^{j+1}$, but we do not know the interval of time between $\mathbf{s}_i^j$ and $\mathbf{s}_i^{j+1}$). The first m out of N samples of each user serve for computing its biometric reference. The remaining ones are used as test samples (for both intraclass verification and interclass attacks). For each user i, two sets of scores are computed. The intra-scores where each test sample of user i is compared to its own biometric reference $\mathbf{r}_i$ ($intra_i = d(\mathbf{s}_i^j, \mathbf{r}_i), \forall j \in [m+1; N]$). The inter-scores where the reference of each user i is tested against the test samples of the other users. ($inter_i = d(\mathbf{s}_l^j, \mathbf{r}_i), \forall j \in [m+1; N], \forall l, l \neq i$). $d(\mathbf{sample}, \mathbf{reference})$ computes the distance between the sample and the biometric reference. Each score is annotated by the number (order of creation) of the test sample. To catch the correlation between scores and capture order, we use the absolute value of the Pearson correlation factor $\rho_{pearson}(\mathbf{X}, \mathbf{Y})$ and the Spearman's rank correlation coefficient $\rho_{spearman}(\mathbf{X}, \mathbf{Y})$. $\mathbf{X}$ represents the scores, while $\mathbf{Y}$ represents the associated capture number.

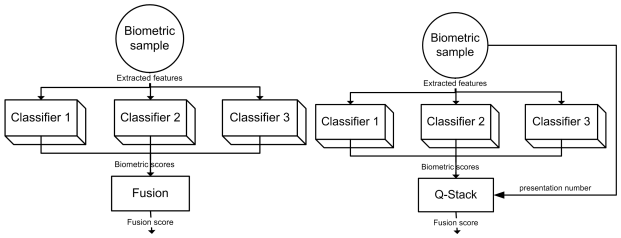
C. Stacked classification

After being able to correlate scores with time, it is interesting to create a classifier which uses the creation number in consideration. To this end, a Q-stack classifier is used in [3]: a first classifier computes the recognition score, while a second classifier uses this score and a time information to accept or reject the query. The aim of the stacked classifier is to take into account the score deviation with time. In this work, we use such classifier with each keystroke dynamics authentication method. We have also modified the stacked classifier so that it can work with a multimodal scheme. In this case, the stacked classifier takes in input: the comparison scores of each keystroke dynamics authentication method, and, the creation number of the query sample[1]. The classifier is a Support Vector Machine [12] which has been trained with samples of the first half of users of the database, and tested with the second one (there is no common users in each database). Data is normalized with the *z-score* technique. A three folds cross validation scheme is used to select the best parameters of the SVM (linear kernel with C between 10 and 1000, and RBF kernel with γ between $1e-3$ and $1e-5$). Raw scores are used to compute the performance without stacked classification.

D. Sample weighting

Another idea to take automatically into account the biometric sample ageing is to use a classifier giving less weight to older samples than newer ones. This way more confidence is given to the recent samples. The principle is similar to the stacked classification, but instead of using the query number as a feature, we use it as a weight for the SVM[2].

[1]when removing the creation number, it is a simple score fusion



(a) Normal fusion (b) Q-stack fusion

Fig. 1. Authentication fusion process.

E. User Classification

Users may behave differently on the evolving of their typing pattern. Thus, the way of taking into account this evolution must be different. To assert this point, we have classified the users in three different classes:

- Users having no correlation between time and recognition score (absolute value of the coefficient lower than 0.15). The way of typing does not evolve with time or is not stable. We call it group 1.
- Users having a very small correlation (absolute value of the coefficient lower than 0.3). We call it group 2.
- Users having more or less correlation. We call it group 3.

User correlation is the mean of correlations of each authentication method. Each dataset has been separated on three distinct sub-datasets containing only members of one group. The Stacked classification experiment has been then reproduced for each group in order to observe differences between them.

F. Discussion

Fig. 1(a) and 1(b) respectively presents the authentication scheme with a normal and Q-stacked fusion. The authentication procedure is the same as the normal case for the sample weighting method. Same information is used for training the SVM (except that the samples are weighted by their presentation number in the weighted case).

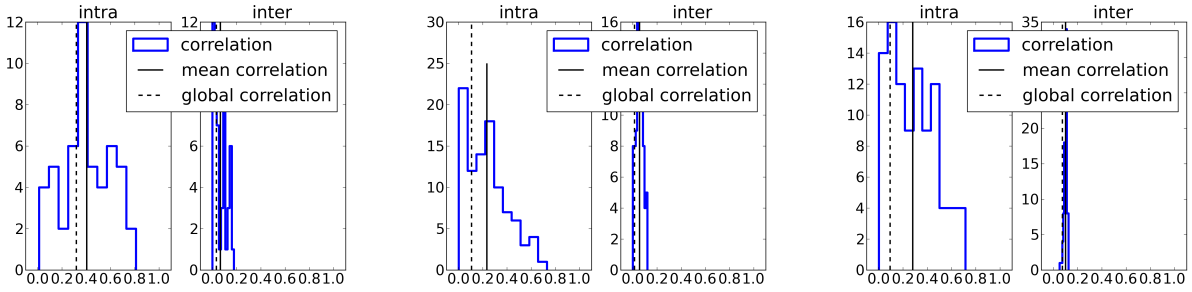
III. Experimental Results

This section presents the obtained results. We first present the result of the analysis of the impact of the chronology when evaluating keystroke dynamics template updating methods. Next, we present the results when using a stacked classifier for such problematic. Then, we present a classification of users.

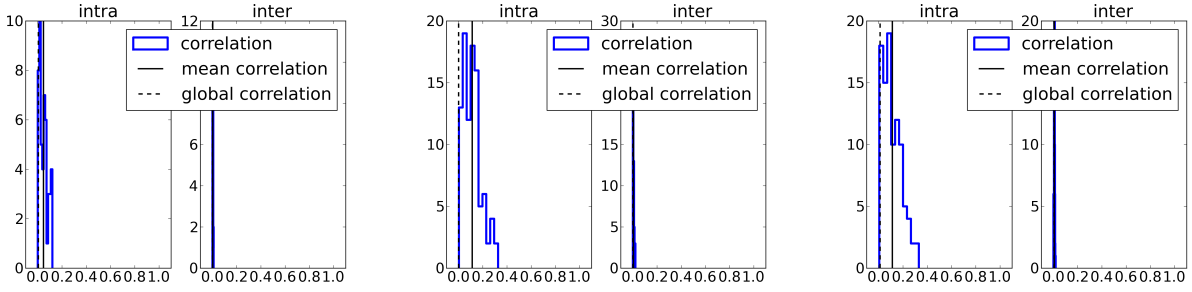
A. Evolution of Scores Through Time

Figure 2 presents the correlation histogram between the recognition method's score and the creation order of the query using the Pearson correlation coefficient (similar results are obtained with the Spearman's rank correlation coefficient). The results are similar for the other keystroke dynamics recognition

[2]This is a feature of scikit-learn which is based on libSVM



(a) chronology/DSL2009/method [2] (b) chronology/GREYC/method [2] (c) chronology/GREYC/rhythm



(d) random/DSL2009/method [2] (e) random/GREYC/method [2] (f) random/GREYC/rhythm

Fig. 2. Correlation histogram for the keystroke dynamics recognition methods when respecting and not chronology

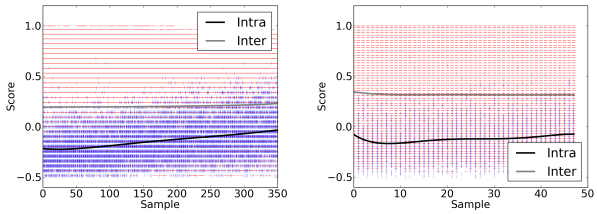
methods and the other database[3]. "Correlation" represents the histogram of the correlation value of the whole set of users. "Mean correlation" is the mean of all the correlation values. "Global correlation" is the correlation when using the whole set of couples of score and presentation number of all the users (we are mainly interested by this value). The correlation has been computed by keeping the chronological order of the capture (*i.e.*, capture with presentation number n have been captured at the n position) and by shuffling the samples order (*i.e.* capture with presentation number n have been captured at the m position, with n which may be different than m). We can do the following assertions:

- The correlation is more important for the intra-scores than the inter-scores, in all the schemes.
- There is far less correlations when shuffling the samples than when respecting it.

Globally, the capture number has an impact on the scores: there is always more correlation when the chronology is respected. Unfortunately, even if we can notice there is some correlation, it is not enough important in term of Person coefficient to be meaningful (we should expect values closer to one to conclude on real correlation). By the way the correlation can be not very important for a lot of users in all the schemes.

The correlation is verified on figure 3, which represents (for all the users) the comparison score depending on the presentation number, and the regression line of these scores. Results are similar for both databases. As expected, the regression lines for the inter-scores are quite horizontal, while

[3]Due to a lack of space, we do not present all combinations



(a) DSL2009/Method [2] (b) GREYC/Method [2]

Fig. 3. Score distribution over time (red for inter-scores and blue for intra-scores). Grey line is the regression line for the inter-scores, while black line is the regression line for the intra-scores.

the regression lines for the intra-scores has higher values with higher presentation number. The big overlap on the intra and inter-scores distribution renders the problem quite difficult, and explains the loss of performance over time.

B. Stacked Classification

Table I presents the performance of the keystroke dynamics methods on the two databases. We compare the performances at the Equal Error Rate (EER). Each method has been tested without ("Normal") and with ("Stacked") the stacked classifier. As expected, scores fusion give better results than each monomodal system. Performances on the DSL2009 datasets are worst than performances on the GREYC dataset, because:

- the password is more complex for DSL2009, and
- its number of sessions is more important.

The use of a stacked classifier does not always improve the results in the case of keystroke dynamics (in opposition to the

TABLE I
PERFORMANCE OF THE TESTED KEYSTROKE DYNAMICS METHODS WHEN USING TIME INFORMATION. CONFIDENCE INTERVALS OF EER AT 90% ARE PRESENTED AS WELL AS THE IMPROVEMENT OF THE STACKED VERSION AGAINST THE CLASSICAL ONE.

	GREYC [5]			DSL2009 [9]		
	Normal	Stacked	Gain (%)	Normal	Stacked	Gain (%)
[6]	[17.39-17.91]	[8.84-26.53]	-0.20	[32.19-32.53]	[17.10-51.30]	-5.68
[2]	[15.56-16.04]	[7.85-23.54]	0.71	[21.80-22.10]	[10.88-32.64]	0.85
Rhythm	[33.72-34.37]	[16.98-50.95]	0.23	[33.85-34.20]	[17.02-51.05]	-0.04
SVM	[14.30-14.78]	[7.38-22.13]	-1.46	[22.04-22.35]	[10.91-32.74]	1.66
WSUM	[15.77-16.27]	[8.04-24.11]	-0.31	[20.62-20.92]	[10.41-31.22]	-0.21
WSVM	[14.27-15.16]			[20.58-21.07]		

TABLE II
PERCENTAGE OF MEMBERS PER GROUP IN EACH DATASET.

Dataset	Group 1	Group 2	Group 3
GREYC	49.48%	29.90%	20.62%
DSL2009	27.45%	27.45%	45.10%

TABLE III
PERFORMANCE ON THE THREE GROUPS OF USER.

	GREYC [5]			DSL2009 [9]		
	Normal	Stacked	Gain (%)	Normal	Stacked	Gain (%)
SVM/gr. 1	[14.12-14.85]	[7.14-21.42]	1.43	[27.47-28.19]	[15.75-47.26]	-13.22
SVM/gr. 2	[13.03-13.98]	[6.68-20.04]	1.06	[15.45-16.05]	[7.77-23.32]	1.28
SVM/gr. 3	[14.83-16.11]	[8.23-24.68]	-6.36	[38.31-38.70]	[20.35-61.05]	-5.69

face recognition problem), which is confirmed by using the Kruskall-Wallis test on the EERs. Performance are improved for some users and decreased for some other ones (thus the EER confidence interval is far wider).

C. User Classification

Table II presents the percentage of user presents in each group, for each dataset. The group distribution is totally different for the two datasets. This can be explained by the differences in the acquisition procedure (complex password, more captures in shorter period, and more sessions for DSL2009). Table III presents the performance using a classic SVM fusion and the Q-stack classifier on the three groups. In all the cases, the systems perform better for users of group 2 (*i.e.*, small correlation). Worst performance is obtained with users of group 3 even with the Q-stack classifier which is supposed to prevent this case.

D. Discussion

We have observed that there is a relation between the intra-scores and the presentation number of the query for some users. This relation explains why performances decrease during the use of keystroke dynamics systems. But, using a stacked classifier to bring this information in the classifier does not always improve the performance (on the contrary to a similar study for face recognition [3]). This may be explained by the fact that the evolution of the biometric data is not as smooth as for the face, and therefore a model re-estimation may be the best alternative. Another explanation is that such system may be only be beneficial for user having the greatest correlation value between score and presentation number.

One limit of our study is that we did not take into account the timestamp of the different biometric samples, but, only their order of capture. This may explain why scores are not as much correlated as in [3], where the authors know the time difference (in days) between several samples.

IV. CONCLUSION

We have studied the correlation between the comparison score of keystroke dynamics authentication methods, and, the chronological number of the query. We have observed that the scores are slightly correlated to the number of the samples (which can be interpreted as the number of times the user has used the system, and then learned the action of typing a password). This correlation justifies the interest of using stacked classifiers (as for face recognition systems). Contrarily to what we expected, we did not observe in our study a real improvement while using a stacked classifier. This can be explained by the fact that: these biometric data do not evolve linearly over time, and, for each user, the evolution is done differently (remember that we use different users to train the stacked classifier and to test it). We can therefore claim that doing a kind of decision frontier evolution over time is less efficient than applying template update procedures (for behavioral systems, when action learning must be taken into account) [4]. Future work will consider using the same protocol with other keystroke dynamics methods (maybe less efficient methods on short term will give better results on long term) while using also model re-estimation, and collecting a database providing more data on a longer timespan.

REFERENCES

[1] L. Araujo, J. Sucupira, L.H.R., M. Lizarraga, L. Ling, and J. Yabu-Uti. User authentication through typing biometrics features. *IEEE Transactions on Signal Processing*, 53:851–855, 2005.

[2] T. de Magalhaes, K. Revett, and H. Santos. Password secured sites: stepping forward with keystroke dynamics. In *International Conference on Next Generation Web Services Practices*, 2005.

[3] A. Drygajlo, W. Li, and K. Zhu. Q-stack aging model for face verification. In *Proc. 17th European Signal Processing Conference (EUSIPCO 2009)*, 2009.

[4] R. Giot, B. Dorizzi, and C. Rosenberger. Analysis of template update strategies for keystroke dynamics. In *IEEE Symposium Series in Computational Intelligence 2011 (SSCI 2011). CIBIM workshop*, 2011.

[5] R. Giot, M. El-Abed, and R. Christophe. Greyc keystroke: a benchmark for keystroke dynamics biometric systems. In *IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS 2009)*, 2009.

[6] S. Hocquet, J.-Y. Ramel, and H. Cardot. Fusion of methods for keystroke dynamic authentication. In *Automatic Identification Advanced Technologies, 2005. Fourth IEEE Workshop on*, 2005.

[7] N. Houmani, S. Garcia-Salicetti, and B. Dorizzi. On assessing the robustness of pen coordinates, pen pressure and pen inclination to time variability with personal entropy. In *Biometrics: Theory, Applications, and Systems, 2009. BTAS'09. IEEE 3rd International Conference on*, 2009.

[8] P. Kang, S.-s. Hwang, and S. Cho. Continual retraining of keystroke dynamics based authenticator. In *Proceedings of ICB 2007*, 2007.

[9] K. S. Killourhy and R. A. Maxion. Comparing anomaly-detection algorithms for keystroke dynamics. In *39th Annual International Conference on Dependable Systems and Networks (DSN-2009)*, 2009.

[10] A. Rattani, G. L. Marcialis, and F. Roli. Boosting gallery representativeness by co-updating face and fingerprint verification systems. 5th Summer School for Advanced Studies on Biometrics for Secure Authentication, 2008.

[11] A. Sethuram, E. Patterson, K. Ricanek, and A. Rawls. Improvements and performance evaluation concerning synthetic age progression and face recognition affected by adult aging. In *ICB 2009*, 2009.

[12] V. Vapnik and V. Vapnik. *Statistical learning theory*. Wiley New York, 1998.