Performance Evaluation of Biometric Template Update

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*—Template update allows to modify the biometric reference of a user while he uses the biometric system. With such kind of mechanism we expect the biometric system uses always an up to date representation of the user, by capturing his intra-class (temporary or permanent) variability. Although several studies exist in the literature, there is no commonly adopted evaluation scheme. This does not ease the comparison of the different systems of the literature. In this paper, we show that using different evaluation procedures can lead in different, and contradictory, interpretations of the results. We use a keystroke dynamics (which is a modality suffering of template ageing quickly) template update system on a dataset consisting of height different sessions to illustrate this point. Even if we do not answer to this problematic, it shows that it is necessary to normalize the template update evaluation procedures.**

***Index Terms*—template update, biometric, evaluation**

I. Introduction

Template update is an active research field whose aim is to update the biometric reference of individuals while using the biometric system. Even if the reason of using template update systems are various (template ageings, noisy acquisitions, lack of samples during enrollment, ...), the expected result is always the same: the improvement of the recognition performance.

Template update mechanisms may vary depending of different factors (which are not directly subject of this work, as we are interested on the evaluation of this mechanism):

- The choice of the of update criteria (threshold, graph based, ...).
- The periodicity of the template update (online and batch, or offline, at various frequencies).
- The working mode of the template update system (supervised or semi-supervised): in the first case, we guaranty no impostor data has been used for the template update.
- The template update mechanism (mainly the employed method used to modify the biometric references).

A very nice work [1], exposes the various points of differences to specify in the studies [1] (they argue that these informations are mainly missing in studies). Nevertheless, this work does not explore the performance evaluation procedure computation (they give information about the way of evaluating the system, but not on the way of computing the error rate). It is necessary to quantify the performance evolution using such kind of mechanism. We will show that different performance computing methods lead to different interpretations of the results. In this work, we present the differences in the various template update (or related) evaluation schemes in the literature. We do not emphasize on the template update mechanisms. We raise the questions that must be answered by the template update community in order to allow an easy evaluation and comparison of the template update mechanisms.

The paper is organised as following. Section II briefly presents the datasets used in the literature for works on template update. Section III presents the different ways encountered in the literature to evaluate the template updating schemes. Section IV illustrates the problem of not having a common evaluation methodology in the template update studies. Section V raises various open questions on template update evaluation methodology.

II. Available Public Databases

Studies on template update require adequate datasets. Various datasets have been used in the literature. They all differ in number of subjects, number of samples per subjects, number of sessions, time difference between the youngest and oldest sample, type of variability. . . The following datasets have been used in the literature in template update works or in studies analysing the variability of samples through time:

- 2D face recognition: there are several datasets for face recognition. In this case, the variabilities are mainly due to pose or illumination differences, but few datasets allow the study of templates ageing by capturing data on a very long period while having a lot of users.
  - The *Equinox Face Dataset* [2] is often used but does not seem to be yet freely available. The number of individuals and samples varies between studies (they do not use the same subset).
  - The dataset *MORPH* [3] has been used in several studies. Once again, the number of individuals and samples varies in studies.
  - The *UMIST Face database* contains 564 images of 20 individuals. Most studies in the state of the art do

[1] although specific to keystroke dynamics

not use the whole set.
  - The *AR* [4] contains several color images of 120 individuals captured on two sessions.
  - Drygajlo *et al.* [5] used youtube's videos of people providing their face each day during three years in general. The timespan is superior to the other datasets, but the number of users is very low and no ground truth is available (automatic image extraction can be erroneous, nothing proves that pictures are presented in chronological order,...).
  - *VADANA* [6] is the most recent dataset designed especially for template update in face recognition systems. 43 subjects have in average 53 pictures, delta between two pictures of an individual can be of several years. This dataset has more intra-class comparison than other long term datasets.
- 3D face recognition. The *Face Recognition Grand Challenge (FRGC) Experiment 3* [7] provides 3D faces linked to color information. Dataset is splitted in a training set of 270 individuals and a testing in of 410 individuals.
- Fingerprint recognition. The dataset [8] comes from the competition "Fingerprint Verification Competition". Four different sub-datasets are available. Each of them contains 110 fingers with 8 samples per finger. This dataset is not appropriate to study variation through time, but it is interesting because of the high intra-class variability of users [9].
- Keystroke Dynamics.
  - The *GREYC keystroke* [10] dataset has been captured among 5 distinct sessions with 100 individuals.
  - The *DSN2009* [11] has been captured amoung 8 distinct sessions with 51 individuals.
- Handwritten signature. The dataset *MCYT-100* [12] is a multimodal biometric database (fingerprint and handwritten signature) which has been used to verify the reliably of extracted features through time [13].

We can see there are various datasets available for several different biometric modalities ; they are summarised in the table I. Most dataset are related to 2D face recognition which is a morphological modality which hold less variability than any behavioral biometric. The properties of these datasets are really different. Few of them have been captured in a long timespan. They are more useful to analyze the intra-class variability due to temporary variations than template ageing.

In the next section, we present the existing evaluation schemes for template update algorithms.

TABLE I
SUMMARY OF THE DATASETS USED IN THE LITERATURE. FIGURES ARE RELATED TO STUDIES USING THE DATASET AND MAY BE DIFFERENT FROM THE REAL VALUE OF THE DATASET. WE CAN SEE THAN FEW OF THEM SEEM APPROPRIATE FOR TEMPLATE UPDATE STUDIES.

Database	# users	# samples	# sessions
2D face			
EQUINOX	40-50	20-100	-
MORPH	14	> 20	-
UMIST	20	25-55	-
AR	120	26	2
YOUTUBE videos	4	1200	1200
VADANA	43	≈53	-
3D face			
FRGC-EXP3	410+270	1-22	-
Fingerprint			
FVC2002	110	8	1
Keystroke dynamics			
GREYC2009	100	60	5
DSN2009	51	400	8
Handwritten signature			
MCYT-100	100	25	5

III. Existing Evaluation Schemes

Few template update studies exist in the literature. In this section, we present the different evaluation protocols found in the literature, using datasets separated in several sessions (also called *batch* in some studies), or not. We also present the different ways of presenting the queries to the biometric references.

A. Studies With Several Sessions

Using dataset providing several capture sessions allows computing error rates specific to sessions. This way, we can track the evolution of the template update through time. Curiously, it is only recently that this kind of evaluation has been encountered [14], [15]. Maybe, such kind of studies is not common because the data acquisition is not very straightforward and too much time consuming.

In such kind of studies, the first session is used to compute the biometric reference of each user, while the next ones are used to apply the template update mechanism and evaluate the update procedure. We can observe two main evaluation processes:

- An *online* order where the comparison score of the query against the reference is used to compute the evaluation measure (and is not only used in the template update mechanism).
- An *offline* order where the comparison score of the query against the reference is not used to compute the evaluation measure. When the whole query set of the session is consumed, the entire query set of the next session is used to evaluate the new biometric references. Following this step, this set is then used for the template update procedure.

Our personal investigations suspect that these two evaluation schemes do not give fundamentally different results, and that the online scheme must be favored to the offline one because:

1) it simplifies the evaluation procedure,
2) it avoids unnecessary computations,
3) it produces an additional session result (as the latest session does not need an additional session to be evaluated).

We have also met two different ways of presenting the results:

- One performance measure per session [14] computed

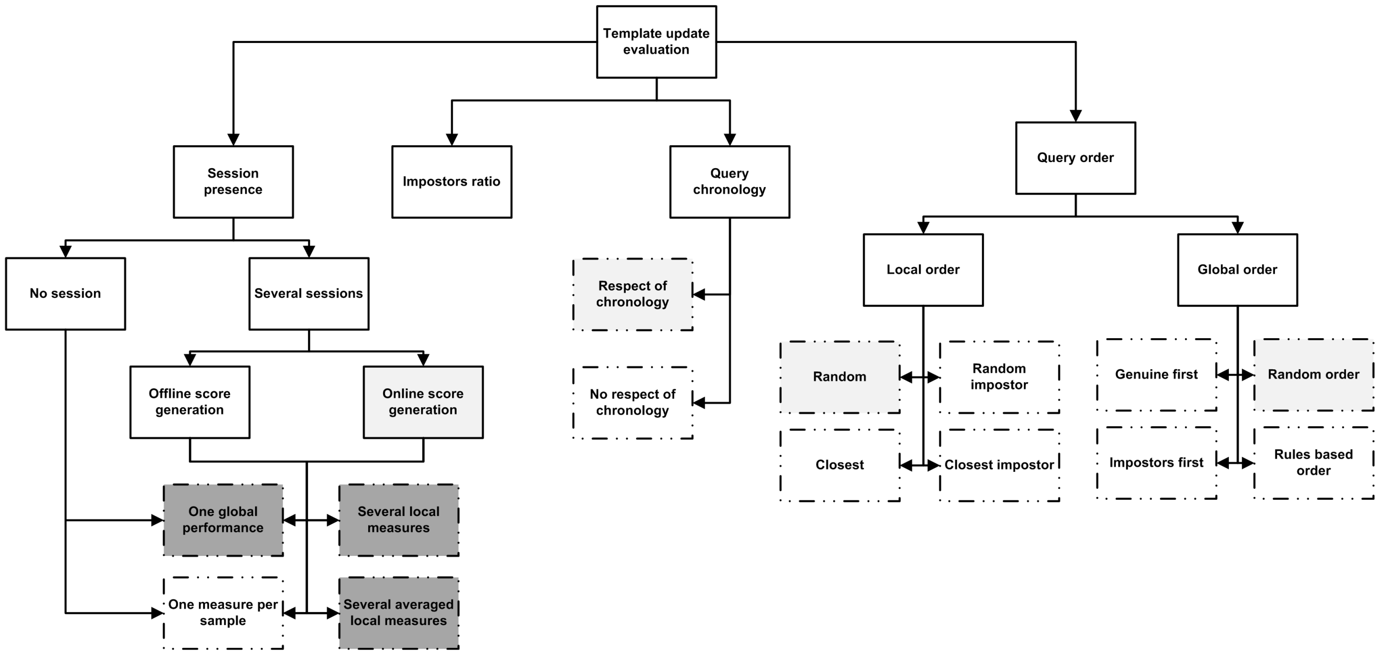


Fig. 1. Summary of all the possible variabilities in a template update evaluation. Dotted nodes represent the possible configuration values, while nodes with a straight line represent the configuration types. Dark gray nodes represent the variant factors in Section IV, while light gray nodes represent the fixed factors in Section IV.

with one of the previously presented methods. This gives result specific to each sample of the session.

- One performance measure per session computed by averaging the performance of the current session and the previous ones [15]. Authors argue this is important because the error rate depends too much on the used test. This smoothing reduces the error rates in comparison to the previous method.
- One global performance computed with the whole set of scores [1].

B. Studies Without Any Session

Most template update studies use datasets with no session, but samples captured in a more or less long period. We can observe two main evaluation procedures:

- Separation of the dataset in two (or three) sub-datasets, which act as if they were two sessions dependant datasets. In this case, the applied procedures are similar to the previously presented ones [16]. In this case, we have only one performance measure for the template update system on the entire dataset.
- Computation of the biometric performance at any time, by modeling its behavior [17]. Note that this method has been illustrated in order to observe the behaviour of a biometric system using no template update system. But, we think it can be used in order to evaluate the performance of an online template update system.

C. Query Presentation Order

Another factor, in the template studies, is the query samples presentation order. We think this information belongs to the evaluation procedure and not directly the template update system, because performance is dependent of them.

In [1], authors make the distinction between global and local orders.

1) Global order: The differences can be:

- The proportion of impostor samples: this is a very important information, as this factor highly impacts the performance: many impostor samples increases the probability of including impostor samples in the biometric references and decreases the performance. This information may be unavailable, fixed at one specific value (50% for example), or several ratios can be specified [14].
- The presentation order of the different types (genuine or impostor) of samples. This is also an important information, as this factor can also impact the performance by driving the probability of doing wrong template updates. We mainly meet three different behaviors. Depending on the studies, one [14], [15] or all [18] of them can be present. The behaviors are:
  - Presenting the *genuine samples first*. All the genuine samples are presented before the impostor samples. Before presenting the first impostor query, the biometric reference might already be highly specialised to efficiently recognize genuine queries and reject impostor queries. We expect really good recognition rates and few impostor samples inclusion in the biometric reference.
  - Presenting the *impostor samples first*. All the impostor samples are presented before the genuine samples. Before presenting the first genuine query, the biometric reference migh already be highly un-

specialised and performs poor results (by having included too many impostor samples and no genuine ones to counterbalance that). We expect quite poor recognition rates and a lot of impostor samples inclusion in the biometric reference.
  - *Random order presentation*. No specific order is preferred. The presentation order is totally random (although controlled by the impostor ratio). A good template update system should include a lot of genuine samples and few impostor samples, while a bad template system includes a lot of impostor samples and few genuine samples. Performances are averaged but probably more realistic than in the first two cases. Of course, this must be done for different impostor ratios.
  - *Rules based order*. The order is directed by a set of rules to follow. Such kind of order is problem specific.

2) Local Order: The local order pays attention to the order of presenting impostors samples.

- *Totally random*. A random sample from a random impostor is selected.
- *Closest*. The closest sample (among all the samples of all the impostors) from the biometric reference is chosen.
- *Random impostor*. An impostor is chosen randomly. His samples are used, in a chronological order for behavioral biometrics, until another impostor is selected.
- *Closest impostor*. The impostor closer to the biometric reference is selected. His samples are used, in a chronological order for behavioral biometrics, until another impostor is selected.

D. Query Chronology

The last important information, regarding the evaluation, is the respect, or not, to the chronology information. When this information is presented, we met two kinds of papers:

- *No chronology respect*. In these papers, samples chronology is not respected. It means that a query B tested against a biometric reference after a query A can be younger than A. In average:

$$\mathbb{P}(age(A) < age(B)) = \mathbb{P}(age(B) < age(A)) \quad (1)$$

  with $\mathbb{P}(e)$ the probability of the event e and $age(s)$ the age of the sample s. This procedure is the most common in the literature whereas it can only be efficient if we assume that the template variability is not related to ageing but other factors. This is of course false for the behavioral modalities and not always true for the morphological ones.
- *Respect of the chronology*. The assumption is that biometric sample variability is also related to ageing of the biometric data (whatever the reason). Genuine samples are always presented by chronological order, but not necessary impostor samples:

$$\mathbb{P}(age(A) < age(B)) = 1 \quad (2)$$

$$\mathbb{P}(age(A) \geq age(B)) = 0 \quad (3)$$

From this review of the literature, we observe that all studies use different protocols, and, that up to now, no standard evaluation procedure exits. Figure 1 summarised the various points subject of variations. It could not be a problem if all these points are indicated in studies [1], because they can be representative of different but useful scenarios. However, when the performance evaluation procedure differs, it can hold to no similar results.

We will illustrate the problem that such a situation can provide in Section IV.

IV. Illustration

The previous section presents the various differences in the evaluation procedure of a template update system. The variation of one factor holds to another testing scenario. We have not discussed about the evaluation of these scenarios.

In this example, we are interested in the evaluation of a template update mechanism [14] for a keystroke dynamics [19] system using the Equal Error Rate (EER) as the evaluation metric. We are not interested in the characteristics of the template update system. This system, which is presented in [14] aims at applying a semi-supervised update based on an update threshold. We have selected two different configurations of the template update system:

- *System 1*: a scenario where the update threshold (distances can be negative) is -0.2.
- *System 2*: a scenario where the update threshold is -0.3.

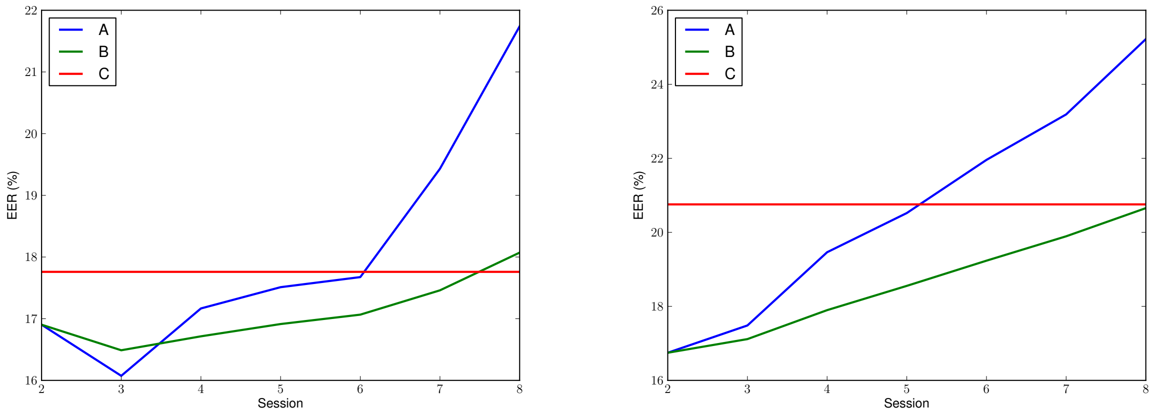
The following fixed parameters are used for the evaluation:

- The dataset [11] provides *8 sessions*. The ways of computing the performance measure are presented later. The first session serves to compute the initial biometric reference. The other sessions serve to update the reference and compute the performance of the updating system.
- We compute the scores for each session in an *online* way.
- The *impostor ratio* is 30%.
- As it is a behavioral modality, we *respect the chronology*.
- The global order of presentation of genuine or impostor samples is *random*.
- The local order of presentation of impostor samples is random too.

This configuration allows us to compute the comparison scores while the system is updating. In addition of these fixed parameters, we have chosen to select three different ways of computing the performance value from these comparison scores. Three different evaluation procedures are applied (the selected performance indice is the EER):

- *Performance evaluation A*. As done in our previous work [14] where the scores of the current session are used to compute its performance.

$$A_i = \text{EER}\left(scores_i\right), \quad \forall i, 2 \leq i \leq S \quad (4)$$



(a) Template update system 1

(b) Template update system 2

Fig. 2. Performances depending on the evaluation method on the same score set. (for the definition of A, B, C see Section IV).

with S the number of sessions, EER$(\cdot)$ the EER computing function and $scores_i$ the scores computed at session i (intra and inter comparisons). We have one EER per session.

$$\mathbf{A} = [A_2, \ldots, A_S] \tag{5}$$

- *Performance evaluation B*. As done in [15] where performance of current session is computed by the mean of all the previous session performance (including the current one).

$$B_i = \frac{1}{i-1} \sum_{j=2}^{i} \text{EER}\left(scores_j\right), \quad \forall i, 2 \leq i \leq S \tag{6}$$

We also have one EER per session.

$$\mathbf{B} = [B_2, \ldots, B_S] \tag{7}$$

- *Performance evaluation C*. As done in [1] where only one measure is computed. In the present case, we merge all the scores of all the sessions in one global set and compute the performance measure on this set.

$$C = \text{EER}\left(\bigcup_{i=2}^{S} scores_i\right) \tag{8}$$

We have one EER for the whole interval. To compare it easily with the two other methods we duplicate it the number of test sessions times.

$$\mathbf{C} = [\underbrace{C, \ldots, C}_{S-1 \text{ values}}] \tag{9}$$

This evaluation procedure is repeated ten times and the results are averaged (as the process is stochastic due to the impostor choices and order).

Figure 1 presents in light gray this fixed configurations and in dark gray the varying configurations. Figure 2 presents the performance, on exactly the same set of scores, of the three evaluation schemes A, B and C. Although globally, the three different evaluations show that system 1 is better than system 2 (better update involving lower EER), we can propose totally different interpretations of the updating system, depending on the chosen evaluation scheme:

- *Performance evaluation A*. Performance of system A decreases fast with time, the template update system does not perform well. The template update system must be improved, or the biometric modality has a very low permanence.
- *Performance evaluation B*. Performance of system B decreases with time, but the amount of decreases is not really important, the template update system is not too bad. The template update is not perfect (there is a performance decrease) but it takes quite well the ageing into account.
- *Performance evaluation C*. Performance of system C is averaged, but we cannot know if it is because of template ageing, because of a bad algorithm or because of a bad dataset.

As no performance measure of a system without template update is presented, we cannot compare the template update systems against the baseline classifier. By the way, the performance evaluation of a system without template update would hold the same performance evaluation problem. The performance evaluation C brings less information than the two other ones. So it must be avoided, because we lack the temporal information which is the most important one. However performance evaluation A and performance evaluation B track temporal evaluation, but give different interpretations. Which one is the most interresting or accurate? In the next section, we raise the questions it would be interesting to answer in order to normalize template update evaluation.

V. Open Questions

All along this paper, we have analyzed the differences in the evaluation protocols, one can encounter in the various biometric template update studies. The variability found in all the protocols raise many open questions:

- What are the characteristics of an interesting dataset for such kind of studies? We have seen that there are several datasets available for the different modalities; they are different in their sample distribution. Few of them seem really interesting to be used in template update scenarios. It is important to know what are the interesting characteristics to respect in order to create new useful datasets.
- What is the best evaluation procedure in order to easily compare the systems without doing each time all the previous experiments from scratch ? The update evaluation procedure is not yet standardized and procedures are really different between studies. Maybe, it is interesting to create new metrics specific for such kind of problem. Some studies present the ratio of impostors included in the updated biometric reference, but other metrics could be interesting too.
- Is it more informative to work with datasets separated in several sessions, or with datasets captured in a longer period without more information ? We can suspect that:
  - In the first case, we have datasets with a small intra-class variability within sessions and a bigger variability between sessions.
  - In the second case, we have datasets with in an intra-class variability homogeneously spread other time.

Without answering these questions, it will be hard to homogenize and compare the different studies on template update mechanisms.

VI. Conclusion

We have presented the different template update evaluation schemes encountered in the literature. We can observe that there exist lots of different and incompatible ways to do it. This hardly allows the comparison of template update mechanisms and their understanding. This asserts the request for the researchers of being very accurate while explaining the experimental protocol in order to ease the reproducibility of the experiment.

References

[1] M. M. Seeger and P. Bours, "How to comprehensively describe a biometric update mechanisms for keystroke authentication," in *3rd International Workshop on Security and Communication Networks. IWSCN*, 2011, pp. 1–7.

[2] "Equinox corporation." [Online]. Available: http://www.equinoxsensors.com/products/HID.html

[3] K. Ricanek and T. Tesafaye, "Morph: A longitudinal image database of normal adult age-progression," in *7th International Conference on Automatic Face and Gesture Recognition. FGR.*, 2006, pp. 341–345.

[4] A. Martinez and R. Benavente, "The ar face database," CVC Technical report, Tech. Rep., 1998.

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

[6] R. Gowri Somanath and C. Kambhamettu, "Vadana: A dense dataset for facial image analysis," in *BeFIT 2011 – First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies*, 2011.

[7] P. Phillips, P. Flynn, T. Scruggs, K. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR.*, vol. 1, 2005, pp. 947–954.

[8] "Fvc2002," 2002. [Online]. Available: http://bias.csr.unibo.it/fvc2002/

[9] B. Freni, G. Marcialis, and F. Roli, "Replacement algorithms for fingerprint template update," in *International Conference on Image Analysis and Recognition. ICIAR*. Springer, 2008, pp. 884–893.

[10] R. Giot, M. El-Abed, and C. Rosenberger, "Greyc keystroke: a benchmark for keystroke dynamics biometric systems," in *IEEE International Conference on Biometrics: Theory, Applications and Systems. BTAS*, Sep. 2009, pp. 1–6. [Online]. Available: http://hal.archives-ouvertes.fr/hal-00432768/en/

[11] K. Killourhy and R. Maxion, "Comparing anomaly-detection algorithms for keystroke dynamics," in *IEEE/IFIP International Conference on Dependable Systems & Networks. DSN*, 2009, pp. 125–134.

[12] J. Ortega-Garcia, J. Fierrez-Aguilar, D. Simon, J. Gonzalez, M. Faundez-Zanuy, V. Espinosa, A. Satue, I. Hernaez, J. Igarza, C. Vivaracho *et al.*, "Mcyt baseline corpus: a bimodal biometric database," in *Vision, Image and Signal Processing*, vol. 150, no. 6, 2004, pp. 395–401.

[13] 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 *IEEE International Conference on Biometrics: Theory, Applications, and Systems. BTAS*, 2009.

[14] R. Giot, B. Dorizzi, and C. Rosenberger, "Analysis of template update strategies for keystroke dynamics," in *Workshop on Computational Intelligence in Biometrics and Identity Management (CIBIM). Special Session on Adaptive Classification Systems for Biometric Recognition. IEEE Symposium Series in Computational Intelligence. SSCI.*, Paris, France, Apr. 2011, pp. 21–28. [Online]. Available: http://hal.archives-ouvertes.fr/hal-00587106/

[15] A. Rattani, G. L. Marcialis, and F. Roli, "Self adaptive systems: An experimental analysis of the performance over time," in *Workshop on Computational Intelligence in Biometrics and Identity Management (CIBIM). Special Session on Adaptive Classification Systems for Biometric Recognition. IEEE Symposium Series in Computational Intelligence. SSCI*, 2011.

[16] F. Roli, L. Didaci, and G. L. Marcialis, *Advances in Biometrics*. SpringerLink, 2008, ch. Adaptive Biometric Systems That Can Improve with Use, pp. 447–471.

[17] N. Poh, J. Kittler, R. Smith, and J. Tena, "A method for estimating authentication performance over time, with applications to face biometrics," in *12th Iberoamerican Congress on Pattern Recognition. CIARP*, 2007.

[18] C. Ryu, H. Kim, and A. K. Jain, "Template adaptation based fingerprint verification," in *18th International Conference on Pattern Recognition, 2006. ICPR.*, 2006.

[19] R. Giot, M. El-Abed, and C. Rosenberger, "Keystroke dynamics overview," in *Biometrics / Book 1*, D. J. Yang, Ed. InTech, Jul. 2011, vol. 1, ch. 8, pp. 157–182. [Online]. Available: http://www.intechopen.com/articles/show/title/keystroke-dynamics-overview