Pattern Recognition 45 (2012) 2622–2632

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Synthetic on-line signature generation. Part II: Experimental validation

Javier Galbally [a,*], Julian Fierrez [a], Javier Ortega-Garcia [a], Réjean Plamondon [b]

[a] Biometric Recognition Group—ATVS, EPS, Universidad Autonoma de Madrid, C/Francisco Tomas y Valiente 11, 28049 Madrid, Spain
[b] Laboratoire Scribens, Dép. de Génie Électrique, École Polytechnique de Montréal, 2900 Boulevard Edouard-Montpetit, Montréal, Canada QC H3T 1J4

ARTICLE INFO

Article history:
Received 9 September 2010
Received in revised form
28 October 2011
Accepted 10 December 2011
Available online 19 December 2011

Keywords:
On-line signature
Synthetic generation
Spectral analysis
Kinematic theory of rapid human movements
Sigma-lognormal model
Biometric recognition

ABSTRACT

A novel method for the generation of synthetic on-line signatures based on the spectral analysis and the Kinematic Theory of rapid human movements, was presented in Part I of this series of two papers. In the present paper, the experimental framework used for the validation of the novel approach is described. The validation protocol, which uses different development and test sets in order to achieve unbiased results, includes the comparison of real and synthetic databases in terms of (i) visual appearance, (ii) statistical information, and (iii) performance evaluation of three competitive and totally different verification systems. The experimental results show the high similarity existing between synthetically generated and humanly produced samples, and the great potential of the method for the study of the signature trait.

1. Introduction

In Part I of this series of two papers [1], we introduced the theoretical framework of a novel method for the generation of synthetic on-line signatures. The proposed model-based algorithm is capable of producing fully synthetic specimens using the combined information obtained from spectral analysis and the Kinematic Theory of rapid human movements, giving freedom both in the number of subjects and samples per user to be generated. In this way, the methodology constitutes a very effective tool to overcome the usual shortage of biometric data without undertaking highly resource-consuming acquisition campaigns. In the present paper we describe the validation protocol followed in order to determine the degree of similarity existing between real and synthetic databases, and we analyze the experimental results obtained. Hence, the main objective of this second part is to present a consistent and replicable evaluation methodology and results which validate the general on-line signature generation approach presented in the preceding paper.

As it was introduced in Part I of this research work, the challenge of generating artificial biometric samples is twofold: (i) on the one hand, the intrinsic information contained in each of the synthetic impressions has to be similar to that comprised within human-produced traits, and (ii) on the other hand, the generation algorithm must be able to produce, in a fully automatic way, synthetic datasets where the overall performance and behavior of a wide range of biometric recognition systems working on that particular trait is consistent and as close as possible, to that obtained on real databases.

Therefore, when validating an approach for the generation of synthetic biometric traits, the problem to be faced is to determine a way to measure, in a quantitative manner, the *realism* of the synthetically produced samples. That is, to define the set of needs that a synthetic sample has to satisfy in order to be recognized and treated by automatic verification systems as a physically collected trait. For the particular case of on-line signature, we can distinguish three different requirements that should be met by synthetic samples. These requirements are closely related to the twofold challenge of synthetic biometric traits generation exposed above, and the experimental framework presented in Sections 3–5 is focused on giving quantitative measures for each of them.

- **Requirement 1: appearance**: Synthetic signatures should look as close as possible to real signatures (i.e., they should have a signature-like visual appearance). This requirement is difficult to quantify as it partly depends on the subjective evaluation of the observer.
- **Requirement 2: information**: Synthetic signatures should have the same statistical characteristics as real signatures. This statistical information can be divided into: (i) topological properties (related to geometry, direction and pressure), (ii) spectral properties (related to time and frequency), and (iii) kinematical properties (related to the speed and acceleration of the strokes).

* Corresponding author. Tel.: +34 91 497 6210; fax: +34 91 497 2107.
E-mail addresses: javier.galbally@uam.es (J. Galbally), julian.fierrez@uam.es (J. Fierrez), javier.ortega@uam.es (J. Ortega-Garcia), rejean.plamondon@polymtl.ca (R. Plamondon).

0031-3203/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2011.12.007

- **Requirement 3: performance**: Synthetic signature databases should present the same inter- and intra-user variability as real signature datasets, which means that the performance of signature verification systems should be as similar as possible when tested on synthetic and real databases.

The new model-based approach for realistic signature generation proposed in the previous article [1] is conceived to produce samples which largely meet these three requirements.

The experimental framework has been designed to establish the level of compliance of the novel scheme proposed, with the twofold challenge posed by the synthetic traits generation problem. The tests comprise results aimed to evaluate to what extent the synthetic signatures present the same type of information as human produced samples (i.e., challenge (i)), and experiments where the global behavior of signature verification systems is assessed both on real and synthetically generated databases (i.e., challenge (ii)). In all the experimental framework, different datasets have been used in the development and test stages in order to obtain totally unbiased results.

The validation protocol includes the visual comparison of the artificial samples appearance and that of real signatures (requirement 1), and the quantitative comparison between the distributions of different distinctive signature global features in the real and synthetic databases (requirement 2). In the recognition experiments, we compare the performance of three state-of-the-art signature verification systems (working on totally different features and matchers), using two real databases and two synthetic datasets generated following the proposed scheme (requirement 3). The different results obtained show the high degree of similarity existing between the synthetic and real signatures and the suitability of the proposed technique for the automatic generation of fully synthetic on-line signature databases.

As was presented in Part I, the proposed synthetic approach is a general method that can generate, depending on the value of its parameters, databases with different levels of intra- and inter-user variability (i.e., different degrees of *difficulty* in terms of automatic signature recognition). In the present work we have tried to adapt the generation method to the level of variability found in standard real databases of western-European signatures such as MCYT [2] or BiosecurID [3]. Although other measures have been proposed in the literature [4], the most common metric to account for the *difficulty* of a database is the performance evaluation of different verification systems. This is the perspective followed in the third validation experiment described in the present article.

The rest of the article is structured as follows. In Section 2 the development (Section 2.1) and test (Section 2.2) databases used in the validation protocol are presented. The experimental results are divided in: appearance tests, Section 3; information comparisons, Section 4; and performance evaluations, Section 5. Conclusions are finally drawn in Section 6.

2. Databases

In order to avoid biased results, four different datasets are used in the experiments. One development set (comprising western-European real signatures) where the parameters of the generation model are estimated, and three test sets, one real (different from the development set, but also containing western-European signers) and two synthetic, where comparative results on the performance of the generation algorithm are obtained. In Fig. 1 we show the general validation strategy followed to evaluate the methodology for synthetic on-line signature generation described in Part I [1], with the databases used in the development and test stages.

2.1. Development database: BiosecurID

The parameters which define the method for synthetic on-line signature generation proposed in [1] (shown here in Fig. 1) are:

- **Synthetic individuals**: N (signature length), N_R (number of relevant spectral coefficients), G (power ratio), PU (number of penups), S (signature slope), F (round-like flourish length).

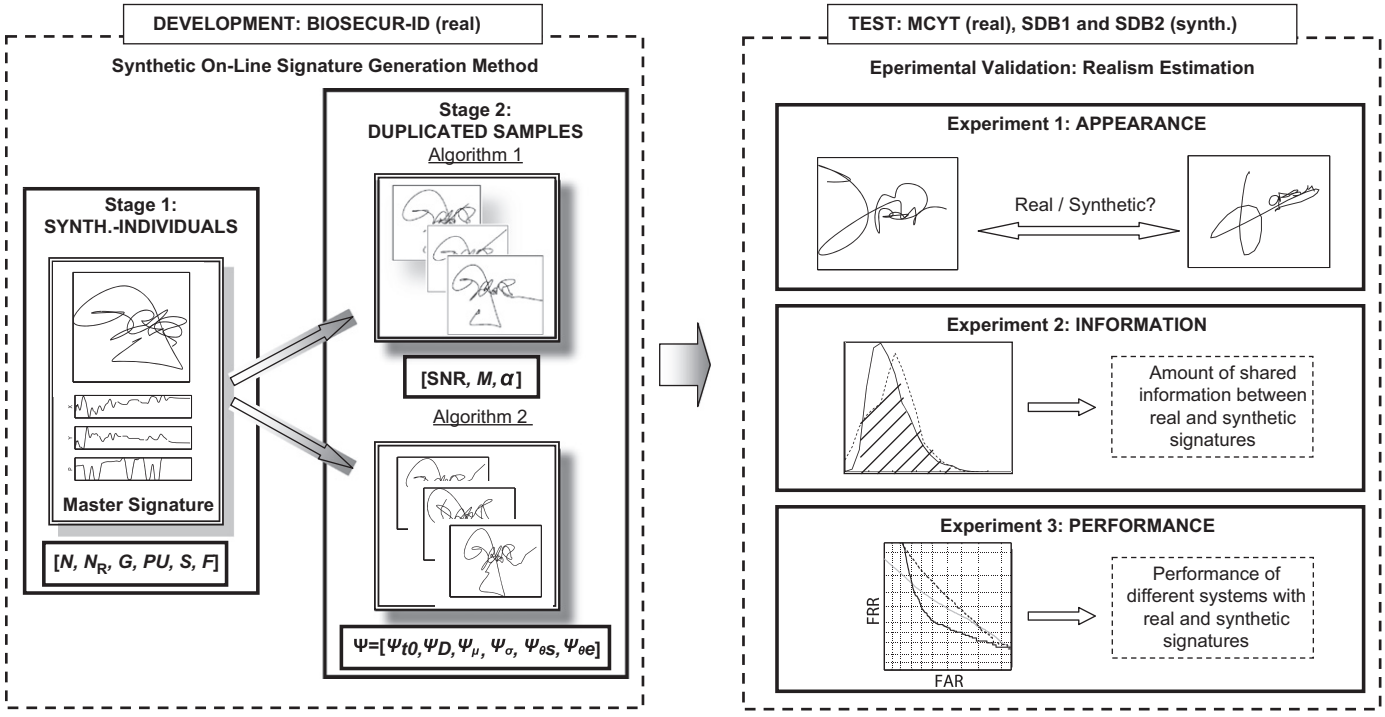

Fig. 1. Validation strategy followed to evaluate the methodology for synthetic on-line signature generation described in Part I [1].

- **Duplicated samples**:
  - **Algorithm 1**: $\mathrm{SNR}=[\mathrm{SNR}_x,\mathrm{SNR}_y]$ (noise addition), M (resampling), and $\alpha=[\alpha_x,\alpha_y,\alpha_p]$ (amplification). With different SNR, M and α for intra-session and inter-session samples.
  - **Algorithm 2**: $\Psi=[\psi_{t_0},\psi_D,\psi_\mu,\psi_\sigma,\psi_{\theta_s},\psi_{\theta_e}]$ (distortion matrix of the Sigma-Lognormal parameters). Again with different values for inter-session and intra-session variability.

For the estimation of these parameters we used the signature data in the BiosecurID multimodal database [3] which was acquired in five different Spanish universities. BiosecurID, comprises eight different biometric traits of 400 users and was captured in four acquisition sessions over a 6-month time span (which makes it an efficient tool to estimate the inter and intra-session variability). The signature subset comprises for each user, 16 original samples (four samples per session), and 12 forgeries carried out with an increasing degree of skill over the sessions (both the off-line and on-line information of each signature is available.) In the present work, the imitations were discarded and the $400\times16=6400$ genuine dynamic signatures were used as development set.

The values obtained on this dataset for each of the parameters defining our generation model of synthetic individuals are given in Appendix A.

2.2. Test databases: MCYT, SDB1, and SDB2

Three different databases, one real and two synthetic, are used in the experiments as test data in order to evaluate the performance of the synthetic on-line signature generation scheme described in Part I of the present report:

- **Real test set: MCYT**: The dynamic signature data of the MCYT database (comprising signature and fingerprint information of 330 users) is used as real test set [2]. This way we ensure that no overlap exists with the development set: different acquisition protocols and sites, different users, number of samples, etc. The signature subcorpus in MCYT is formed by 25 original samples and 25 skilled forgeries per user (captured in five different acquisition sets). For the validation experiments the original data are used while the forged samples are discarded.
- **Synthetic test sets: SDB1 and SDB2**: Two different synthetic databases were produced for the experiments, one using Algorithm 1 for the generation of duplicated samples (SDB1), and the other following Algorithm 2 (SDB2). Both synthetic datasets follow the MCYT structure, comprising 330 different signers with 25 samples per user. The first 5 of those 25 signatures are generated using the intra-session values of the model parameters (estimated from the development set BiosecurID), and the remaining 20 specimens present a higher variability in order to imitate samples acquired in different sessions (inter-session values of the parameters).

Three different experiments were carried out on the previous databases in order to estimate the level of compliance of the synthetic samples with each of the three requirements that should be met by synthetic signatures and which were defined in Section 1.

3. Validation experiment 1: appearance comparison

This first experiment is designed to evaluate from a statistical point of view the subjective perception that non-expert human observers have of synthetic signatures. For this purpose, the set of 100 real and synthetic samples shown in Fig. 2 was given to a group of 25 people with naive knowledge on signature recognition and

Fig. 2. Set of 100 real and synthetic samples used in Experiment 1 to evaluate the realism of the visual appearance of synthetic signatures. Real signatures are highlighted in gray.

they were asked to mark each specimen from 0 (fully synthetic) to 4 (somewhat synthetic) and from 6 (somewhat real) to 10 (fully real) according to their impression after a quick inspection of the signature. The maximum time permitted to complete the experiment was 20 min.

Although sometimes the synthetic generation method produces isolated characters, in general it is not capable of generating handwriting or real names, thus, in order to make the task more difficult and fair, the 50 real signatures were chosen from the MCYT DB with the only restriction that no easily readable name could be distinguished. The 50 synthetic signatures were randomly selected from SDB1 and SDB2 (half from each dataset).

Two types of errors can be committed in the classification task: (i) a real signature is marked as synthetic (0–4), measured by the False Synthetic Rate (FSR), and (ii) a synthetic signature is mistaken

Table 1
Error rates, average scoring and average time produced by the 25 participants in Experiment 1: classifying as real or synthetic the set of 100 signatures shown in Fig. 2. FSR stands for False Synthetic Rate, FRR for False Real Rate, and ACE for Average Classification Error.

Error rates (%)			Average scoring		Average time (min)
FSR	FRR	ACE	Real	Synthetic	
35.28	36.72	36.00	5.91	4.16	11.5

Table 2
Information divergence between real and synthetic databases generated following the methodology described in the present work. Smaller values indicate a higher amount of shared information.

DBs	Relative entropy					
	Time	Direct.	Speed	Geom.	**Total**	20-Best
MCYT–Bio.ID	0.03	0.14	0.05	0.12	**0.09**	0.11
MCYT–SDB1	1.01	0.52	0.30	0.74	**0.64**	0.62
MCYT–SDB2	1.21	0.64	0.90	0.90	**0.92**	1.13
MCYT–Unif.	1.31	1.42	2.93	2.01	**1.92**	1.75

with a real sample (ranked 6–10), measured by the False Real Rate (FRR). The final Average Classification Error (ACE) is defined as $ACE = (FSR + FRR)/2$. These error rates are presented in the first three columns of Table 1. In the next two columns we give the average scoring given by all 25 subjects to the 50 real and synthetic samples. Finally the average time taken to complete the experiment is shown.

From the results presented in Table 1 we can see that over one third of the signatures (36%) were misclassified, proving the real-like appearance of synthetic samples (a random classifier would present an ACE of 50%). It should also be noticed that both error rates FSR and FRR are very similar (35.28% and 36.72%, respectively) which means that the number of mistaken real and synthetic samples is very similar and that it is not easier to distinguish one class over the other.

As well, the FSR obtained on the samples of both synthetic databases, SDB1 and SDB2, was almost identical (33.66% and 36.90%, respectively), showing that the visual appearance of the samples generated by either method is very similar and close to that of real signatures.

Furthermore, the average scoring given by the users to real (5.91) and synthetic specimens (4.16) is quite close, reinforcing the idea that human subjects have a very similar perception of both types of signatures.

Finally, we have to remark that the average time taken by the users to carry out the experiment was 11.5 min (around 7 s per signature), which is fully aligned with the overall objective of the experiment of not making a detailed and profound analysis of each signature, but estimating the general visual appearance of synthetic samples after a short inspection.

4. Validation experiment 2: information estimation

In addition to the observable similarity between the real and synthetic signatures appearance (patent from the results obtained in the first experiment), two other experiments were carried out in order to assess the suitability of the proposed synthetic signature generation algorithm.

In this second experiment, we evaluated the compliance of the synthetic generation algorithm with requirement 2. With this objective, we studied to what extent the synthetic signatures in SDB1 and SDB2 present the same information as the real signatures in MCYT, according to the comprehensive set of 100 global features described in [5].

In the experiment, this 100-feature set, which comprises many of the features of the most popular works on feature-based signature verification [6], is extracted from each signature in MCYT and from both synthetic databases. Then, in order to give a measure of the common information, the individual distributions of each parameter for real and synthetic samples are computed and the similarity between both types of signatures (real and synthetic) is estimated as the Kullback–Leibler divergence [7] (also named relative entropy or information divergence).

The Kullback–Leibler divergence $D(R\|S)$ is used in probability theory and information theory as a way to measure the difference between two probability distributions R and S. Typically R represents the *true* distribution of data, observations, or a precise calculated theoretical distribution which, in this particular case, will be represented by each parameter distribution extracted from MCYT, R_i with $i = 1 \ldots 100$. The distribution S typically represents a theory, model, description, or approximation of R which in this case will be each parameter distribution extracted from SDB1 and SDB2, S_i with $i = 1 \ldots 100$.

For probability distributions R_i and S_i of a discrete random variable X their K–L divergence is defined as

$$D_i(R_i\|S_i) = \sum_{x \in \mathcal{X}} R_i(x) \log \frac{R_i(x)}{S_i(x)}. \quad (1)$$

The K–L divergence for the whole 100 feature set is then computed as $D = 1/100 \sum_i D_i$. The relative entropy is always nonnegative and is zero if and only if $R = S$, that is, the smaller the K–L divergence the higher the shared information by the two distributions R and S.

The parameters comprised in the 100 feature set considered in this experiment can be classified according to the signature property measured in [8]: (i) time related features, (ii) direction related features, (iii) speed and acceleration related features, and (iv) geometry related features. The amount of information present in synthetic signatures of SDB1 and SDB2 for each of these groups according to the K–L metric defined above (Eq. (1)) is given in the first four columns of Table 2, while the fifth column corresponds to the information for the whole feature set. As the distributions of the set of parameters may also vary among databases comprising real signatures, for completion and also as a baseline result, in Table 2 the information divergence in the two real databases used in the validation protocol, BiosecurID and MCYT, is also given. Also for reference, the value of the relative entropy between the parameter distributions extracted from MCYT and uniform distributions is given in the last row of Table 2.

Apart from the previous classification, the 100-feature set has been studied in signature verification tasks [9], where a best performing 20-parameter subset was found using the Sequential Forward Floating Selection (SFFS) algorithm [10]. The K–L divergence between real and synthetic signatures for this subset is given in the last column of Table 2.

Several observations can be extracted from the results presented in Table 2: (i) the K–L divergence between real databases for all the groups of parameters analyzed is very small and consistent (always around 0.1), which suggests that the 100-feature set is a good way to condense the information contained in signatures; (ii) the amount of shared information between real and synthetic datasets is quite big (K–L divergence lower than 0.7 for the best synthetic dataset), and for all cases clearly higher than the similarity obtained with a uniform distribution; (iii) the information divergence is slightly lower for all subsets considered in the case of SDB1 compared to SDB2 (0.64

against 0.92 when the whole 100 feature set is taken into account). From this last observation, we can infer that the direct modification of the Sigma–Lognormal parameters (Algorithm 2 for the generation of duplicated samples) is slightly worse than Algorithm 1 with respect to maintaining the information present in real signatures. This may be explained by the fact that while Algorithm 1 directly modifies the x and y functions, Algorithm 2 first reconstructs the velocity signal, introduces some deformations in it, and then decomposes it again into the new coordinate sequences. In spite of being very accurate, in the latter case, the reconstruction and decomposition processes may introduce some unwanted noise into the x and y functions that can account for the small loss of information observed.

However, Algorithm 2 provides more natural results in some parts of the signatures (e.g. it better preserves the nonlinearity of the time fluctuations between samples). This is because Algorithm 2 is based on information extracted from the neuromuscular impulses involved in human handwriting encompassed in the Sigma–Lognormal model. As a result, we can conclude that Algorithm 1 and Algorithm 2 are complementary.

In order to supply also with a visual comparison between distributions in addition to the quantitative measure, the real (solid) and synthetic (dashed) individual distributions for each of the parameters comprised in the best performing 20-feature subset are shown in Fig. 3 (the parameter numeration followed is the same used in [5]). The complete best performing 20-feature subset is given in Appendix B. Although the synthetic distributions have been obtained using SDB1, those corresponding to SDB2 do not present any significant difference to the ones shown here, and are just omitted for clarity.

From the resulting individual distributions of real and synthetic signatures shown in Fig. 3, we can observe the clear correlation that exists between them, not only in the quantitative values, but also in shape and appearance, being in some cases (parameters 1, 8, 21, 34, 57, and 77) practically identical.

We can conclude from the results shown in Table 2 and Fig. 3 that most of the features, and therefore most of the information, that characterize the signature trait, are present in a very similar fashion both in the real and synthetic signatures generated according to the proposed approach.

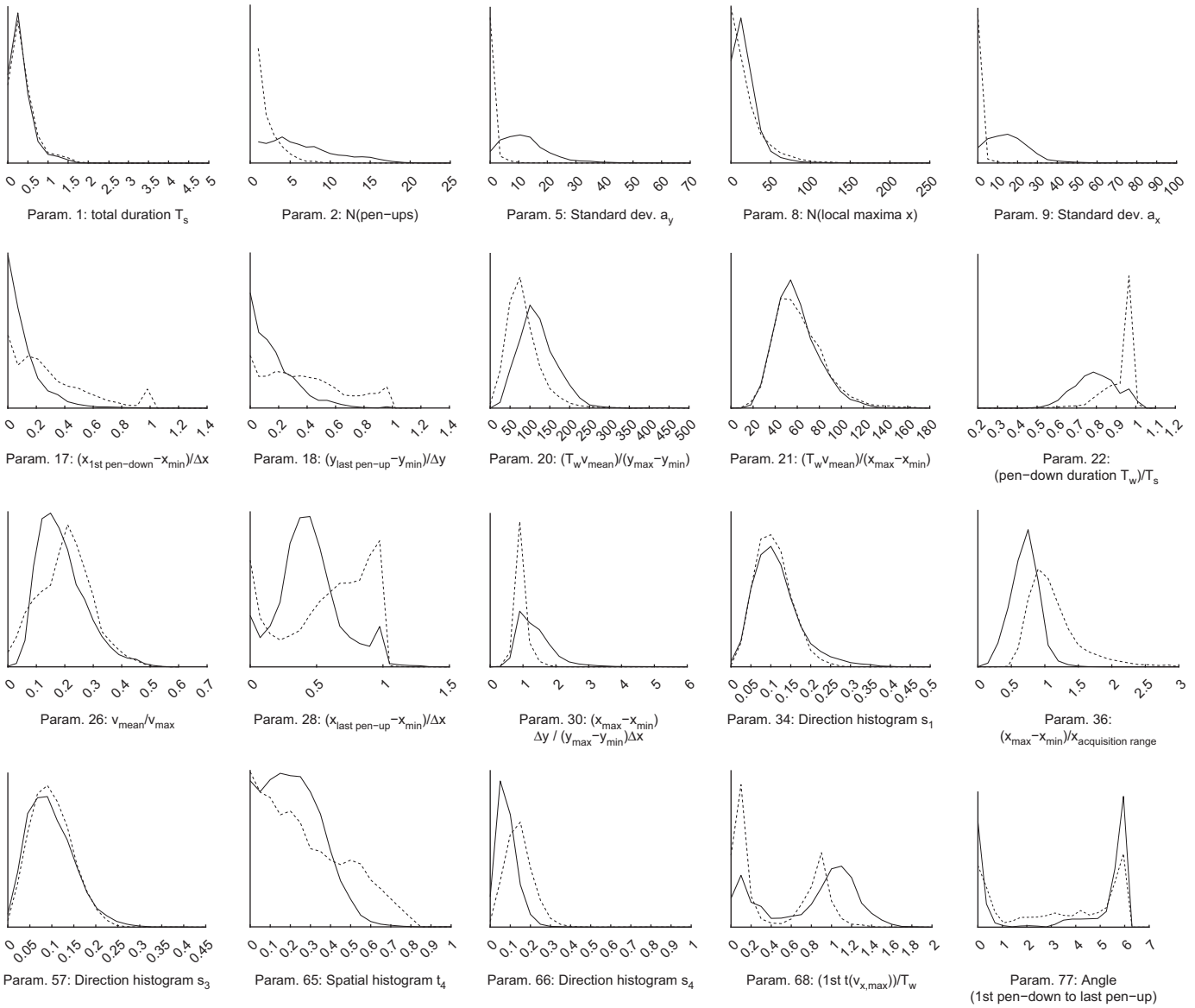

Fig. 3. Histograms of real (solid lines) and synthetic (dashed lines) signatures taken from MCYT and SDB1, corresponding to the best performing 20-parameter set for signature verification [9]. The parameter numeration followed in [5] is used, where the complete set of 100 parameters from which the best 20 were selected was introduced and discussed. The complete list of this 20-parameter set is given in Appendix B.

5. Experiment 3: performance measure

In this last experiment, we focused on the analysis of the third requirement given in Section 1 to be complied by synthetic signatures, that is, if the behavior of signature verification systems is similar when they are evaluated on real and synthetic databases. For this purpose, we have assessed the performance of three different competitive on-line signature verification systems using totally diverse feature sets (feature- and function-based) and matchers (distance measure, Hidden Markov Models, and Dynamic Time Warping), over the three test databases (MCYT, SDB1 and SDB2).

Two different scenarios have been considered in the experiments, namely: (i) a realistic working scenario where a reduced number of samples of each user are available to train its model, and (ii) a hypothetical case study in which we may have many training samples for each user. The protocol followed to compute the set of genuine and impostor scores in each of the cases was:

- *Few training samples*: The first five signatures were used to train the user model, and the remaining 20 samples were used as test set, thus producing $20 \times 330 = 6600$ genuine scores.
- *Many training samples*: The user model was trained with 20 signatures, and the remaining five samples were used as test set, which gives $5 \times 330 = 1650$ genuine scores.

In both cases the set of impostor scores was computed using one signature of the remaining users, which means that we have a total $329 \times 330 = 108,570$ non-genuine scores.

The three on-line verification systems evaluated in the experiments were:

- **System A: feature-based + Mahalanobis distance**: This system models the signature as a holistic multidimensional vector composed of the best performing 40-feature subset extracted in [9] from the total set of 100 global features described in [5] (the analogue best 20-feature subset was already used in experiment 2 of the present work). In the present study, we used this 40-feature representation of the signatures normalizing each of them to the range [0,1] using the tanh-estimators described in [11]. Finally, the similarity scores are computed using the Mahalanobis distance between the input vector and a statistical model of the attacked client estimated using a number of training signatures (few/many depending on the scenario).
- **System B: function-based + HMM**: This function-based verification system applies a regional approach using a statistical model built using Hidden Markov Models (HMMs) [12] to a set of 10 time sequences selected applying the SFFS feature selection algorithm to the total set of 34 functions defined in [13]. This subset of 10 signals are derived from the coordinate (x and y) and pressure (p) functions, while no pen inclination signals are used as its utility for automatic signature recognition is at least unclear [14]. After some preprocessing (position and rotation alignment), and the computation of extended functions (path angle, velocity, curvature, acceleration, and time derivatives) to complete a set of 23 time sequences, similarities are computed using 12 left-to-right HMM states and mixtures of four Gaussians per state. This system participated in the Signature Verification Competition 2004 with very good results [15], and the general configuration is detailed in [16].
- **System C: function-based + DTW**: In this function-based local approach a subset of nine time functions (selected using SFFS from the total 34 feature set as in the case of system B) are directly matched using the elastic technique Dynamic Time Warping (DTW) [17]. Dynamic Time Warping is an application of Dynamic Programming to the problem of matching time sequences of different lengths, thus, the goal of DTW is to find an elastic match among samples of a pair of sequences that minimize a given distance measure. In this particular implementation, which is thoroughly described in [18], we use the Euclidean distance as the measure to be optimized and only three correspondences among samples of the compared sequences are allowed, using symmetrical weighting factors. Although the DTW algorithm has been replaced by more powerful ones such as HMMs or SVMs for speech applications, it remains a highly effective tool for signature verification as it is best suited for small amounts of training data, which is a common case in signature verification.

In the particular context of this experiment, the ultimate goal of biometric traits synthesis would be to produce synthetic databases such that the DET (Detection Error Trade-off) curves obtained on any verification system are as similar as possible to those achieved using real datasets. We cannot forget that one given verification system will not present exactly the same behavior even when evaluated with two different real datasets, thus, a certain variability on the performance among real and synthetically produced data would be not only acceptable but desirable.

The performance results (DET curves) obtained for verification systems A, B and C, following the described experimental protocol, and for the three mentioned databases (MCYT, SDB1 and SDB2), are shown in Fig. 4. We can observe that the curves of the three systems under the two considered scenarios present a very high degree of resemblance, both from a quantitative (EERs) and qualitative (general behavior) point of view, for the case of real and synthetic signatures. Note for example the high similarity of the DET curves for system A with five training signatures.

The results obtained for both synthetic databases are quite remarkable. We may argue that, from a qualitative point of view, SDB1 presents in general a slightly better fit with the performance obtained for MCYT than SDB2. However, if we analyze the results upon the basis of the EER, SDB2 reaches more similar values to MCYT in three out of the six cases, and in two out of three when only the more realistic 5 Tr. scenario is considered (three training signatures is the upper limit for many commercial applications).

The performance of systems B and C on both scenarios (few/many training signatures) was also analyzed without considering the pressure function (which is not captured by all on-line signature acquisition devices). System A was not included in this case since many of its features are based on the pressure function. With this experiment, we have been able to study the impact of including the pressure information in the overall performance of the systems, and study if the synthetic pressure function has a similar effect than the real pressure on the error rates of the verification applications considered. Results are shown in Fig. 5 where we can observe that the performance of the systems worsens compared to the case in which the pressure signal is included. It is important to notice that this increase in the error rates occurs in a very similar way for the case of real and synthetic signatures, which suggests that the artificial pressure function contributes to the general performance of the verification systems in an analogue manner than the humanly produced signal.

Again, when comparing the performance of both synthetic databases, we may observe the same trends as in the scenario where the pressure function was considered: although the subjective appearance of the SDB1 curves is slightly more similar to MCYT than SDB2, the EER reached with SDB2 is closer to that of MCYT for all four tests. Thus, in general, Algorithm 1 or 2 can be

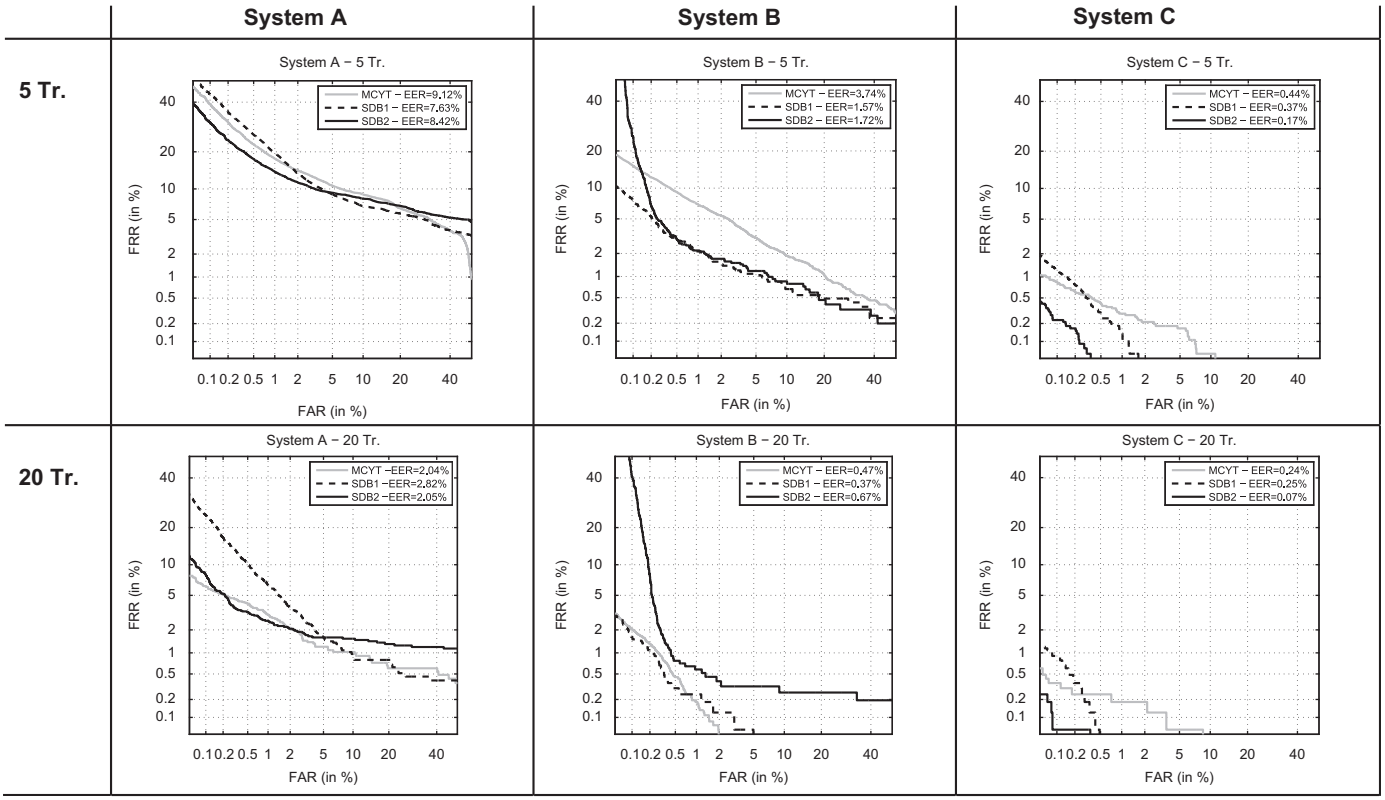

Fig. 4. Comparison of the performance of systems A, B, and C, for 5 and 20 training signatures, on a real (MCYT, gray DET curve) and synthetic databases (SDB1 and SDB2, solid and dashed black DET curves). The EER (Equal Error Rate) is indicated in each plot. FAR stands for False Acceptance Rate, FRR for False Rejection Rate and Tr for training signatures.

used depending on whether the user wants to obtain a better estimation of a system's overall behavior (SDB1), or a more exact quantitative measure of the EER operating point (SDB2).

Moreover, for most cases in both scenarios (with and without taking into account the pressure information), SDB1 provides an upper limit to the real signatures curves, while, in turn, SDB2 produces a lower bound. This may account for the complementarity of both algorithms, which should be jointly used for a better estimation of the performance of on-line signature verification systems.

We may conclude that the results and conclusions derived from this third validation experiment, confirm the great potential of the described generation methodology and prove its suitability for obtaining reliable estimations on the performance of signature verification systems.

6. Conclusions and discussion

The novel method for the generation of synthetic on-line handwritten signatures introduced in the first paper of this series of two papers has been evaluated in the present paper using two different development and test sets in order to avoid biased or over optimistic results. Although several general indications as how to evaluate synthetic databases are given, and some of the ideas proposed and used here may be applicable to the evaluation of synthetic datasets containing other biometric traits, addressing the problem of synthetic database evaluation from a general perspective would constitute in itself a whole new work that falls out of the scope of this research.

The validation protocol included three different experiments where synthetic and real signatures were compared in terms of: (i) visual appearance, (ii) statistical information which they present, and (iii) performance evaluation of three competitive and totally different signature verification systems. In all the tests, the synthetic signatures obtained remarkable results, showing a very high degree of similarity in all the considered scenarios with humanly produced samples.

A comparative evaluation of both duplicated samples generation algorithms was also conducted in the validation experiments. Both schemes, one of them based on some signal processing simplifications and the other based on biomechanical properties of the human handwriting, reached very good results showing a high degree of complementarity specially for performance evaluation purposes.

The validation protocol and results described in the present work have demonstrated that, from a computer-based recognition point of view, the databases produced following the proposed generation approach are fully representative of the different real signatures that may be found in every day life in a western-European context. This was the primary goal of the novel research work described in Parts I and II. From a human perspective, it is clear that some of the signatures have a more realistic appearance than others as the proposed algorithm is not capable of producing readable names but only, by chance, isolated characters. Even though this was not the primary objective of the project, the first evaluation experiment has clearly shown that the synthetic methodology is specially effective generating human-like signatures consisting of just some sort of flourish.

The validation results described in this work have shown that the novel generation method presented in Part I constitutes a very powerful and useful system with a great potential for many different tasks such as: performance estimation [19], security evaluation in order to test existing biometric solutions against fraudulent attempts [20], individuality studies [21], or for synthetically increasing the amount of enrollment data in order to improve the performance of a given application [22].

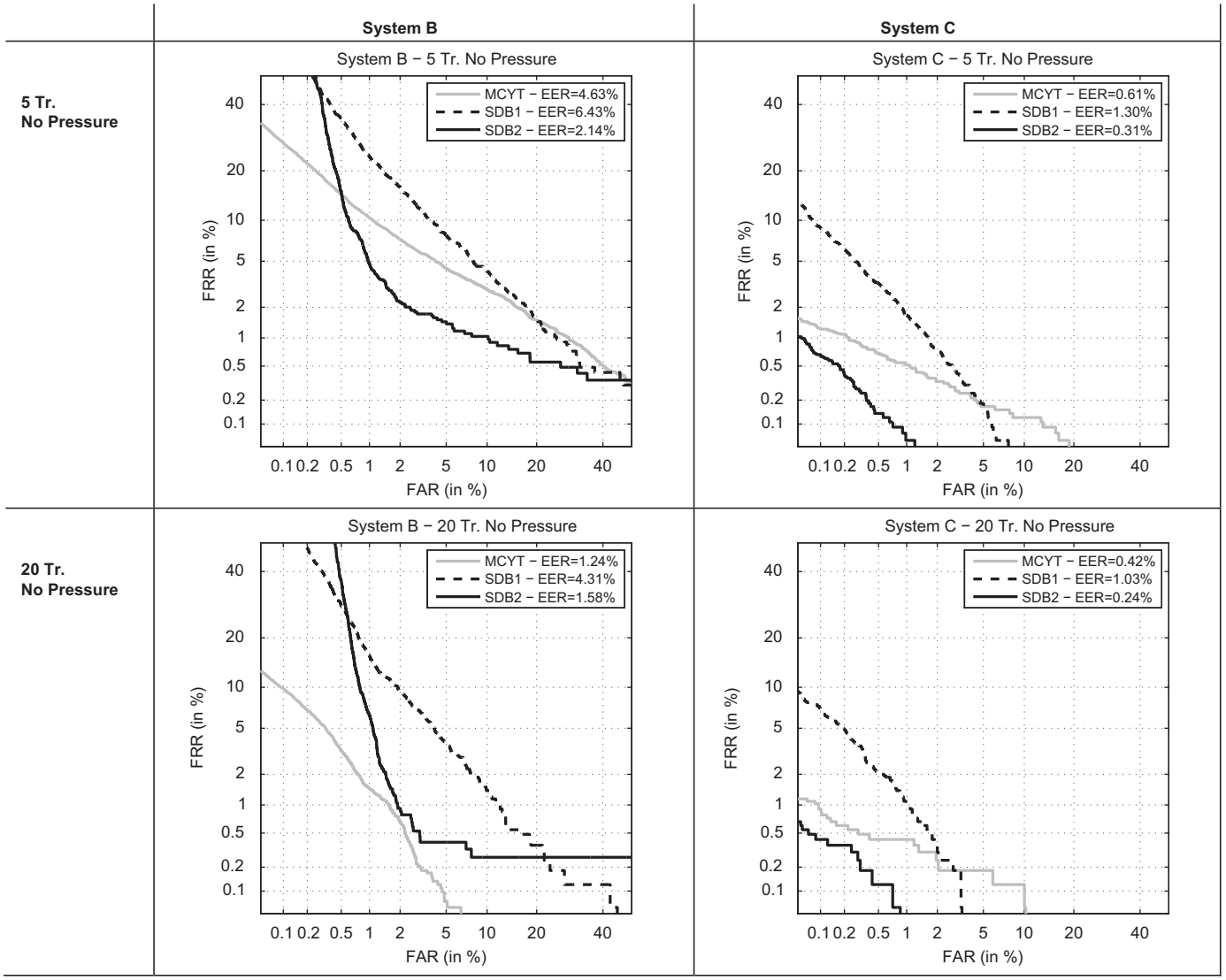

Fig. 5. Comparison of the performance of systems B and C, for 5 and 20 training signatures, on a real (MCYT, gray DET curve) and synthetic databases (SDB1 and SDB2, solid and dashed black DET curves), without taking into account the pressure information. The EER is indicated in each plot. FAR stands for False Acceptance Rate, FRR for False Rejection Rate and Tr for training signatures.

It should be emphasized that the objective of this novel work is not to encourage the substitution of real signatures by synthetic ones, but rather to provide a powerful tool for the development of signature recognition systems. In particular, the synthetic signatures can be used to objectively compare the authentication efficiency, limitations, and capabilities of newly designed verification algorithms through their testing on a large-scale dataset of synthetically generated signatures. However, although it has been proved that synthetic traits contain similar characteristics and information to that of real samples, and therefore constitute a very useful aid for performance estimation, they should not be seen as a substitute but as a complement of real traits and the definitive evaluation of a given system should always be carried out in a realistic working environment and using real data. Therefore, the use of synthetic biometric data should be understood as a valid alternative in order to obtain a fast and reliable estimate of the recognition performance of biometric systems under controlled and repeatable conditions which enable the fair comparison of different algorithms, but in no case as a replacement of human-generated data.

Also to be noticed that the UK Biometrics Working Group [23] has published a set of best practices for testing and reporting performance results of biometric systems [24], where it is advised to avoid adding synthetic data to a test set, or adding noise to the data for scenario testing in order to prevent the bias derived from those practices that often makes the results difficult to interpret. However, neither of these cases is similar to the scenario studied and proposed in this work, where we did not artificially increase the amount of real data with synthetically generated samples, but we created fully synthetic databases on which it was shown that different verification systems present performance results which are consistent with those reached on real datasets.

Acknowledgments

J.G. is supported by a FPU fellowship from Spanish MEC. This work has been partially supported by projects Contexts (S2009/TIC-1485) from CAM, Bio-Challenge (TEC2009-11186) from Spanish MICINN, *irección General de la Guardia Civil* and *átedra UAM-Telefónica*. R.P. is supported by grant RGPIN-915 from NSERC.

J.G. would also like to thank R.P. as the main coordinator of the Scribens Laboratory at the École Polytechnique de Montréal for hosting him during the development of this research work.

Appendix A. Parameter values

The values obtained on the development dataset (BiosecurID) for each of the parameters defining our generation model of synthetic individuals are given below. It should be noticed that using different development and test sets ensures that the feature values shown here are not tuned to obtain over optimistic results in the validation tests.

As the synthetic generation approach is general, these values may be recomputed on different development databases (e.g. containing Asian, North American, or Arabian signatures) in order to produce specimens with more similar characteristics to other type of signers different from the western-European considered in the present work.

The values of the parameters obtained on the development dataset defining the model for the generation of the master signature are:

- **Parameter** N: It follows the length distribution of the development set.
- **Parameter** N_R: The values that define the uniform distribution from which this parameter is extracted are: $[\delta_N^{\min},\delta_N^{\max}]=[0.15,0.26]$, with $N_R=\delta_N N$.
- **Parameter** G: The ratio between the power of the relevant and non-relevant coefficients follows a uniform distribution defined by $G^{\min}=8$ and $G^{\max}=19$.
- **Parameter** PU: It follows the penups distribution of the development set according to the signature length N (i.e., longer signatures present a higher probability of having a bigger number of penups.)
- **Parameter** S: The values that define the uniform distribution which allow to compute the slope of left-to-right written signatures are: $[\delta_S^{\min},\delta^{\max}]=[0.05,0.25]$, with $s_s=\delta_S N$, $s_l=N$, and $S=s_s/s_l$.
- **Parameter** F: The values that defines the uniform distribution from which the length of the round-like flourish is estimated are: $[\delta_F^{\min},\delta_F^{\max}]=[0.08,0.17]$, with $F=\delta_F N$. This waveform is added to the signatures with a probability $p_F=0.37$.

The values of the parameters defining Algorithm 1 (for the generation of duplicated samples) obtained on the development dataset are:

- **Parameter** SNR: Based on the assumption of uncorrelated signature signals and noise, we estimate the SNR averaging the noise (computed between pairs of genuine signatures avoiding repetitions) across users. Thus, the global SNR of signal x of a specific user ($\mathrm{SNR}_x^{\mathrm{U}}$) is estimated as

$$\mathrm{SNR}_x^{\mathrm{U}}=\frac{1}{\mathrm{C}(N_{gs},2)}\sum_{k=1}^{N_{gs}}\frac{P_x^i}{|P_x^i-P_x^j|}\quad \text{for } j>i,$$

where N_{gs} represents the number of considered genuine signatures from the user, and $\mathrm{C}(N_{gs},2)$ is the number of possible combinations of the N_{gs} signatures taken in pairs: $\mathrm{C}(N_{gs},2)=N_{gs}!/2!(N_{gs}-2)!$.

The final SNR_x distribution is estimated using the 400 $\mathrm{SNR}_x^{\mathrm{U}}$ measures obtained from BiosecurID.

Parameter SNR_y is computed similarly, being in both cases the genuine pairs of signatures (N_{gs}) either from the same or different acquisition sessions (intra-session and inter-session SNR models, respectively).

The results show that the power of the noise added in the x coordinate to produce inter-session samples P_{nx}^{inter} has to be around 8% higher than in the case of intra-session repetitions P_{nx}^{intra} (i.e., $P_{nx}^{\mathrm{inter}}=1.08P_{nx}^{\mathrm{intra}}$). In the case of the noise affecting the y coordinate function, the variability between samples captured in the same and different sessions is slightly higher: $P_{ny}^{\mathrm{inter}}=1.11P_{ny}^{\mathrm{intra}}$.

- **Parameter** M: The value of the intra-session duration variability found in the development set is defined by $M^{\mathrm{intra}}=0.1$, while the inter-session variability follows a uniform distribution characterized by $M^{\mathrm{inter}}=0.14$.
- **Parameter** α: The values that define the uniform distributions from which this parameter is extracted are (for the three time functions x, y, and p)

$$[\alpha_x^{\mathrm{intra}},\alpha_x^{\mathrm{inter}}]=[0.06,0.08],$$

$$[\alpha_y^{\mathrm{intra}},\alpha_y^{\mathrm{inter}}]=[0.08,0.11],$$

$$[\alpha_p^{\mathrm{intra}},\alpha_p^{\mathrm{inter}}]=[0.05,0.06].$$

The values of the parameters defining Algorithm 2 (for the generation of duplicated samples) obtained on the development dataset are:

- **Parameter** Ψ: The values that define the uniform distributions from which this parameter is extracted are (for the six Lognormal features)

$$[\psi_{t_0}^{\mathrm{intra}},\psi_{t_0}^{\mathrm{inter}}]=[0.004,0.005],$$

$$[\psi_D^{\mathrm{intra}},\psi_D^{\mathrm{inter}}]=[0.12,0.15],$$

$$[\psi_\mu^{\mathrm{intra}},\psi_\mu^{\mathrm{inter}}]=[0.08,0.1],$$

$$[\psi_\sigma^{\mathrm{intra}},\psi_\sigma^{\mathrm{inter}}]=[0.08,0.1],$$

$$[\psi_{\theta_s}^{\mathrm{intra}},\psi_{\theta_s}^{\mathrm{inter}}]=[0.04,0.06],$$

$$[\psi_{\theta_e}^{\mathrm{intra}},\psi_{\theta_e}^{\mathrm{inter}}]=[0.04,0.06].$$

From these values of Ψ we can see that the most critic Lognormal feature (the one that admits the lowest variation) is t_0, while the most relaxed is D (relatively large variations of this parameter do not change significantly the master signature). The rest of lognormal features (μ, σ, θ_s, and θ_e) accept a similar degree of variation in order to generate realistic duplicated samples following Algorithm 2.

Table 3
Set of best performing 20 global features considered in Section 4 of the present work, sorted following the numeration used in [5] where they were first introduced. T denotes time interval, t denotes time instant, N denotes number of events, and θ denotes angle. Note that all notations are either defined or referenced somewhere in the table (e.g., Δ_x is defined in 17, histograms in 57,65, and 66 are referenced in 34, etc.).

#	Feature description	#	Feature description
1	Signature total duration T_s	**2**	N(pen-ups)
5	Standard deviation of a_y	**8**	N(local maxima in x)
9	Standard deviation of a_x	**17**	$\frac{(x_{1\text{st pen-down}}-x_{\min})}{\Delta_x=\sum_{i=1}^{\text{pen-downs}}(x_{\max(i)}-x_{\min(i)})}$
18	$(y_{\text{last pen-up}}-y_{\min})/\Delta_y$	**20**	(T_w average velocity $\overline{v}$)/($y_{\max}-y_{\min}$)
21	$(T_w\overline{v})/(x_{\max}-x_{\min})$	**22**	(Pen-down duration T_w)/T_s
26	$\overline{v}/v_{\max}$	**28**	$(x_{\text{last pen-up}}-x_{\min})/\Delta_x$
30	$\frac{(x_{\max}-x_{\min})\Delta_y}{(y_{\max}-y_{\min})\Delta_x}$	**34**	Direction histogram s_1 [5]
36	$(x_{\max}-x_{\min})/x_{\text{acquisition range}}$	**57**	Direction histogram s_3
65	Spatial histogram t_4	**66**	Direction histogram s_4
68	(1st $t(v_{x,\max})$)/T_w	**77**	θ(1st pen − down to last pen − up)

Appendix B. Validation experiment 2: 20-parameter set

In Table 3 we show the best performing 20-parameter set found in [9] and used in the validation experiment 2 of the present work to compare the information present in synthetic and real signatures. Each parameter corresponds to the feature distributions shown in Fig. 3. The numeration followed is the same used in [5] where the complete 100-parameter set was first introduced.

References

[1] J. Galbally, R. Plamondon, J. Fierrez, J. Ortega-Garcia, Synthetic on-line signature generation. Part I: methodology and algorithms, Pattern Recognition, doi:10.1016/j.patcog.2011.12.011, this issue.

[2] J. Ortega-Garcia, J. Fierrez, D. Simon, M.F. Gonzalez, V. Espinosa, A. Satue, I. Hernaez, J.J. Igarza, C. Vivaracho, D. Escudero, Q.I. Moro, MCYT baseline corpus: a bimodal biometric database, IEE Proceedings: Vision, Image and Signal Processing 150 (6) (2003) 391–401.

[3] J. Fierrez, J. Galbally, J. Ortega-Garcia, M.R. Freire, F. Alonso-Fernandez, D. Ramos, D.T. Toledano, J. Gonzalez-Rodriguez, J.A. Siguenza, J. Garrido-Salas, E. Anguiano, G.G. de Rivera, R. Ribalda, M. Faundez-Zanuy, J.A. Ortega, V. Cardeñoso-Payo, A. Viloria, C.E. Vivaracho, Q.I. Moro, J.J. Igarza, J. Sanchez, I. Hernaez, C. Orrite-Uruñuela, F. Martinez-Contreras, J.J. Gracia-Roche, BiosecurID: a multimodal biometric database, Pattern Analysis and Applications 13 (2) (2010) 235–246.

[4] S. Garcia, N. Houmani, B. Dorizzi, A client-entropy measure for on-line signatures, in: Proceedings of IEEE Biometrics Symposium (BSym), 2008, pp. 83–88.

[5] J. Fierrez-Aguilar, L. Nanni, J. Lopez-Peñalba, J. Ortega-Garcia, D. Maltoni, An on-line signature verification system based on fusion of local and global information. in: IAPR Audio- and Video-Based Biometric Person Authentication (AVBPA), Lecture Notes in Computer Science, vol. 3546, Springer, 2005, pp. 523–532.

[6] J. Fierrez, J. Ortega-Garcia, On-line signature verification, in: Handbook of Biometrics, Springer, 2008, pp. 189–209.

[7] T.M. Cover, J.A. Thomas, Elements of Information Theory, Wiley-Interscience, 2006.

[8] J. Galbally, J. Fierrez, M.R. Freire, J. Ortega-Garcia, Feature selection based on genetic algorithms for on-line signature verification, in: Proceedings of IEEE Workshop on Automatic Identification Advanced Technologies (AutoID), 2007, pp. 198–203.

[9] J. Galbally, J. Fierrez, J. Ortega-Garcia, Performance and robustness: a trade-off in dynamic signature verification, in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2008, pp. 1697–1700.

[10] P. Pudil, J. Novovicova, J. Kittler, Floating search methods in feature selection, Pattern Recognition Letters 15 (1994) 1119–1125.

[11] A.K. Jain, K. Nandakumar, A. Ross, Score normalization in multimodal biometric systems, Pattern Recognition 38 (2005) 2270–2285.

[12] L.R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proceedings of the IEEE 77 (2) (1989) 257–286.

[13] M. Martinez-Diaz, J. Fierrez, J. Hangai, Signature features, in: Encyclopedia of Biometrics, Springer, 2009, pp. 1185–1192.

[14] J. Fierrez, J. Ortega-Garcia, D. Ramos, J. Gonzalez-Rodriguez, HMM-based on-line signature verification: feature extraction and signature modeling, Pattern Recognition Letters 8 (2007) 2325–2334.

[15] D.Y. Yeung, H. Chang, Y. Xiong, S. George, R. Kashi, T. Matsumoto, G. Rigoll, SVC2004: first international signature verification competition, in: D. Zhang, A.K. Jain (Eds.), Proceedings of IAPR International Conference on Biometrics Authentication (ICBA), Lecture Notes in Computer Science, vol. 3072, Springer, 2004, pp. 16–22.

[16] M. Martinez-Diaz, J. Fierrez, J. Galbally, J. Ortega-Garcia, Towards mobile authentication using dynamic signature verification: useful features and performance evaluation, in: Proceedings of IAPR International Conference on Pattern Recognition (ICPR), 2008.

[17] A. Kholmatov, B. Yanikoglu, Identity authentication using improved online signature verification method, Pattern Recognition Letters 26 (2005) 2400–2408.

[18] M. Martinez-Diaz, J. Fierrez, S. Hangai, Signature matching, in: Encyclopedia of Biometrics, Springer, 2009, pp. 1192–1196.

[19] R. Cappelli, D. Maio, D. Maltoni, J.L. Wayman, A.K. Jain, Performance evaluation of fingerprint verification systems, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (1) (2006) 3–18.

[20] J. Galbally, J. Fierrez, M. Martinez-Diaz, J. Ortega-Garcia, Evaluation of brute-force attack to dynamic signature verification using synthetic samples, in: Proceedings of IAPR International Conference on Document Analysis and Machine Intelligence (ICDAR), 2009.

[21] A. Kholmatov, B. Yanikoglu, An individuality model for online signatures using global Fourier descriptors. in: Proceedings of SPIE Biometric Technology for Human Identification V (BTHI V), vol. 6944, 2008.

[22] J. Galbally, J. Fierrez, M. Martinez-Diaz, J. Ortega-Garcia, Improving the enrollment in dynamic signature verification with synthetic samples, in: Proceedings of IAPR International Conference on Document Analysis and Recognition (ICDAR), 2009.

[23] BWG, Communications-Electronics Security Group (CESG)—Biometric Working Group (BWG) (UK Government), 2009.

[24] A. Mansfield, J. Wayman, Best Practices in Testing and Reporting Performance of Biometric Devices, Technical Report, CESG Biometrics Working Group, 2002.

Javier Galbally received the MSc in Electrical Engineering in 2005 from the Universidad de Cantabria, and the PhD degree in Electrical Engineering in 2009, from Universidad Autonoma de Madrid, Spain. Since 2006 he is with Universidad Autonoma de Madrid, where he is currently working as an assistant researcher. He has carried out different research internships in worldwide leading groups in biometric recognition such as BioLab from Universita di Bologna Italy, IDIAP Research Institute in Switzerland, or the Scribens Laboratory at the École Polytechnique de Montréal in Canada. His research interests are mainly focused on the security evaluation of biometric systems, but also include pattern and biometric recognition, and synthetic generation of biometric traits. He is actively involved in European projects focused on vulnerability assessment of biometrics (e.g., STREP Tabula Rasa) and is the recipient of a number of distinctions, including: IBM Best Student Paper Award at ICPR 2008, finalist of the EBF European Biometric Research Award 2009, and best PhD thesis Award by the Universidad Autonoma de Madrid 2010.

Julian Fierrez received the MSc and the PhD degrees in Electrical Engineering from Universidad Politecnica de Madrid, Madrid, Spain, in 2001 and 2006, respectively. Since 2002 he has been affiliated with the Biometric Recognition Group-ATVS, first at Universidad Politecnica de Madrid, and since 2004 at Universidad Autonoma de Madrid, where he is currently an Associate Professor. From 2007 to 2009 he was a visiting researcher at Michigan State University in USA under a Marie Curie fellowship. In the past, he has also conducted 3-month research visits at Halmstad University in Sweden (2003), Bologna University in Italy (2004), Michigan State University in USA (2005), and University of Surrey in UK (2006). His research interests and areas of expertise include signal and image processing, pattern recognition, and biometrics, with emphasis on signature and fingerprint verification, multibiometrics, biometric databases, and system security. Dr. Fierrez has been and is actively involved in European projects focused on biometrics (e.g. FP6 BioSec IP, FP6 BioSecure NoE, and FP7 BBFor2 Marie Curie ITN), and is the recipient of a number of research distinctions, including: best poster paper at AVBPA 2003, Rosina Ribalta award to the best Spanish PhD proposal in ICT in 2005, best PhD thesis in computer vision and pattern recognition in 2005–2007 by the IAPR Spanish liaison (AERFAI), Motorola best student paper at ICB 2006, EBF European Biometric Industry Award 2006, and IBM best student paper at ICPR 2008.

Javier Ortega-Garcia received the MSc degree in Electrical Engineering (Ingeniero de Telecomunicacion), in 1989; and the PhD degree "cum laude" also in Electrical Engineering (Doctor Ingeniero de Telecomunicacion), in 1996, both from Universidad Politecnica de Madrid, Spain. Dr. Ortega-Garcia is founder and co-director of the Biometric Recognition Group-ATVS. He is currently a Full Professor at the Escuela Politecnica Superior, Universidad Autonoma de Madrid, where he teaches Digital Signal Processing and Biometric Recognition courses. His research interests are focused on biometrics signal processing: speaker recognition, fingerprint recognition, on-line signature verification, data fusion and multibiometrics. He has published over 150 international contributions, including book chapters, refereed journal and conference papers. He chaired "Odyssey-04, The Speaker Recognition Workshop", (co-sponsored by IEEE), and co-chaired "ICB-09, the 3rd IAPR International Conference on Biometrics". He has been appointed as Chair of "ICB-13, the 5th IAPR International Conference on Biometrics".

Réjean Plamondon received a BSc degree in Physics, and MSc and PhD degrees in Electrical Engineering from Université Laval, Québec, P.Q., Canada in 1973, 1975 and 1978 respectively. In 1978, he joined the faculty of the École Polytechnique, Université de Montréal, Montréal, P.Q., Canada, where he is currently a Full Professor. He has been the Head of the Department of Electrical and Computer Engineering from 1996 to 1998 and the Chief Executive Officer of Ecole Polytechnique from 1998 to 2002. He is now the Head of Laboratoire Scribens at this institution. Over the last thirty years, Professor Plamondon has been involved in many pattern recognition projects, particularly in the field of on-line and off-line handwriting analysis and processing. He has proposed many original solutions, based on exhaustive studies of human

movement generation and perception, to problems related to the design of automatic systems for signature verification and handwriting recognition, as well as interactive electronic penpads to help children learning handwriting and powerful methods for analyzing and interpreting neuromuscular signals. His main contribution has been the development of a kinematic theory of rapid human movements which can take into account, with the help of a unique basic equation called a delta-lognormal function, the major psychophysical phenomena reported in studies dealing with rapid movements. The theory has been found successful in describing the basic kinematic properties of velocity profiles as observed in finger, hand, arm, head and eye movements. Professor Plamondon has studied and analyzed these biosignals extensively in order to develop creative and powerful methods and systems in various domains of engineering. Full member of the Canadian Association of Physicists, the Ordre des Ingénieurs du Québec, the Union national des ecrivains du Québec, Dr. Plamondon is an also active member of several international societies. He is a Fellow of the Netherlands Institute for Advanced Study in the Humanities and Social Sciences (NIAS; 1989), of the International Association for Pattern Recognition (IAPR; 1994) and of the Institute of Electrical and Electronics Engineers (IEEE; 2000). From 1990 to 1997, he was the President of the Canadian Image Processing and Pattern Recognition Society and the Canadian representative on the board of Governors of IAPR. He has been the President of the International Graphonomics Society (IGS) from 1995 to 2007. He has been involved in the planning and organization of numerous international conferences and workshops and has worked with scientists from many countries. He is the author or co-author of more than 300 publications and owner of four patents. He has edited or co-edited four books and several Special Issues of scientific journals. He has also published a children book, a novel and three collections of poems.