Synthetic Generation of Handwritten Signatures Based on Spectral Analysis

Javier Galbally, Julian Fierrez, Marcos Martinez-Diaz, and Javier Ortega-Garcia

Biometric Recognition Group–ATVS, EPS, Universidad Autonoma de Madrid
C/ Francisco Tomas y Valiente 11, 28049 Madrid, Spain

ABSTRACT

A new method to generate synthetic online signatures is presented. The algorithm uses a parametrical model to generate the synthetic Discrete Fourier Transform (DFT) of the trajectory signals, which are then refined in the time domain and completed with a synthetic pressure function. Multiple samples of each signature are created so that synthetic databases may be produced. Quantitative and qualitative results are reported, showing that, in addition to presenting a very realistic appearance, the synthetically generated signatures have very similar characteristics to those that enable the recognition of real signatures.

Keywords: Biometrics, synthetic generation, signature verification

1. INTRODUCTION

With the increasing importance that biometric security systems are acquiring in today's society and their introduction in many daily applications, a growing interest is arising for the generation of synthetic biometric traits such as voice,[1] fingerprints,[2] iris,[3] handwriting,[4] or signature.[5] The generation of these synthetic samples is of interest, among other applications, for performance evaluation[6] and vulnerability assessment[7] of biometric systems.

More specifically, synthetically generated biometric databases: *i*) facilitate the performance evaluation of recognition systems instead of the costly and time-consuming real biometric databases, and *ii*) provide a tool with which to evaluate the vulnerability of biometric systems to attacks carried out with synthetically generated traits.

It should be emphasized that, although there are multiple works which address the problem of generating synthetic traits, not all of them consider the term *synthetic* in the same way. In particular, three different strategies for producing synthetic biometric samples can be found in the current literature:

- **Duplicated samples**. In this case the algorithm starts from one or more *real* samples of a given person and, through different transformations, produces different synthetic (or duplicated) samples corresponding to the same person. This type of algorithms are useful to *increase* the amount of already acquired biometric data but not to generate completely new datasets. Therefore, its utility for performance evaluation and vulnerability assessment in biometrics is very limited. On the other hand, this class of methods can be helpful to synthetically augment the size of the enrollment set of signatures in identification and verification systems, a critical parameter in signature biometrics.[8]

  The great majority of existing approaches for synthetic signature generation are based on this type of strategy.[9–13]

- **Combination of different real samples**. This is the approach followed by most speech and handwriting synthesizers. This type of algorithms start from a pool of *real* n-grams (isolated letters, or combination of two or more letters) and using some type of concatenation procedure combine them to form the synthetic samples.[4,14] Again, these techniques present the drawback of needing *real* samples to generate the synthetic trait and therefore their utility for performance evaluation and vulnerability assessment in biometrics is also very limited. As in the previous case, this perspective for the generation of synthetic data is useful to produce multiple biometric samples of a given real user, but not to generate synthetic individuals.

Further author information: javier.galbally@uam.es, Phone: +34 91 497 3363, Fax: +34 91 497 2235

Optics and Photonics in Global Homeland Security V and Biometric Technology for Human Identification VI, edited by Craig S. Halvorson, Šárka O. Southern, B. V. K. Vijaya Kumar, Salil Prabhakar, Arun A. Ross, Proc. of SPIE Vol. 7306, 730629 · © 2009 SPIE · CCC code: 0277-786X/09/$18 · doi: 10.1117/12.817928

- **Synthetic-individuals**. In this case, some kind of *a priori* knowledge about a certain biometric trait (e.g., minutiae distribution, iris structure, signature length, etc.) is used to create a model that characterizes that biometric trait for a population of subjects. New *synthetic* individuals can then be generated sampling the constructed model. In a subsequent stage of the algorithm, multiple samples of the synthetic users can be generated by any of the procedures for creating duplicated samples.

  Regarding performance evaluation and vulnerability assessment in biometrics this approach has the clear advantage over the two previously presented, of not needing any real biometric samples to generate completely synthetic databases. This way, these algorithms constitute a very effective tool to overcome the usual shortage of biometric data without undertaking highly resource-consuming acquisition campaigns.

  Different model-based algorithms have been presented in the literature to generate synthetic individuals for biometric traits such as iris[3] or fingerprint.[2] Bezine *et al.*[15] and Djioua *et al.*[16] proposed two different models to characterize the handwriting process but did not carry out any conclusive experiments regarding the suitability of the models for synthesis purposes. To the best of our knowledge, Popel is the only author who has described this type of approach for synthetic signature generation using a complicated model based on information extracted from the time domain.[5]

In the present article we will describe a new model-based approach for realistic signature generation based on information obtained from the frequency domain, which does not need of any previously acquired real samples. This work studies the synthetic generation of the so called *occidental* signatures. In opposition to other types of signatures consisting of independent symbols, such as the *asian* signatures, the occidental signatures typically consist of handwritten concatenated text and some form of flourish.

The motivation to base our model on spectral analysis comes mainly from two facts. On the one hand, spectral analysis constitutes a general and powerful tool that enables the parameterization of complex time functions such as the ones found in online signature biometrics. This is for example patent in the work of Kholmatov and Yanikoglou,[17] who used it to devise a spectrum-based signature parameterization for their individuality study of the online signature biometrics. On the other hand, working with the spectrum of the signature functions permits us to exploit some similarities that we have heuristically found among different occidental handwritten signatures (this point will be further detailed in Sect. 2).

The validation methodology of the algorithm is based on qualitative and quantitative results which show the suitability of the technique and the high degree of similarity existing between the synthetic signatures generated and real signatures.

The paper is structured as follows. An overview of the algorithm used to generate the synthetic signatures is given in Sect. 2. In Sect. 3 the method used to generate multiple samples of a given signature is presented. The protocol followed in the experiments, together with the results obtained are presented in Sect. 4. Conclusions are finally drawn in Sect. 5.

2. SIGNATURE GENERATION ALGORITHM

Although other signals such as the azimuth and elevation angles of the input pen might be taken into account, in this work we will consider that an online signature is defined by three time sequences $[x[n]\ y[n]\ p[n]]$ specifying each of them the x and y coordinates, and the pressure applied during the signing process at the time instants $n = 1, \ldots, N$ (here sampled at 100 Hz).

The algorithm proposed in the present contribution to generate synthetic signatures comprises three successive steps, as can be seen in Fig. 1. A first step, carried out in the frequency domain, in which the synthetic Discrete Fourier Transform (DFT) of the trajectory signals x and y is generated using a parametrical model, obtained by spectral analysis of a development set of real signatures. In the second stage the resulting trajectory signals are used to place the penups of the pressure function. Finally, in the last stage, all the three signals are processed in the time domain in order to give the synthetic signatures a more realistic appearance. These three steps are described next.

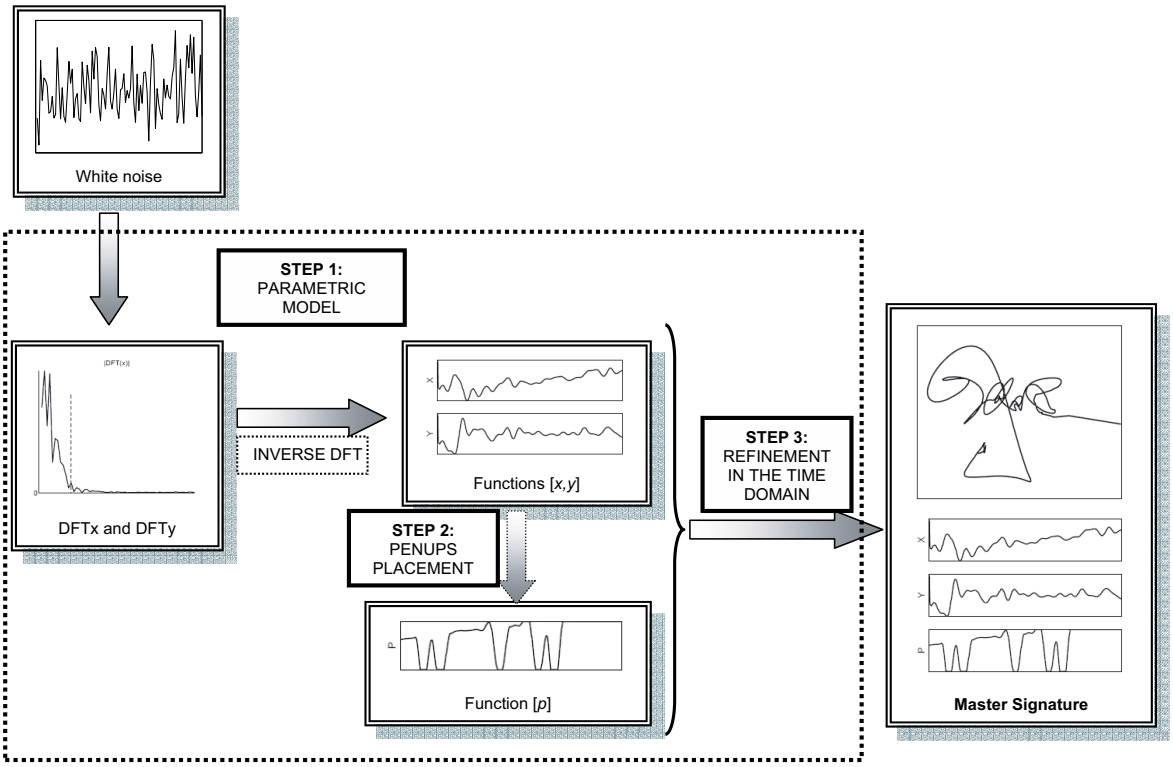


Figure 1. General diagram of the synthetic signature generation algorithm proposed.

2.1 Signature Model in the Frequency Domain

In this step, a parametrical model is used to generate the DFT of the synthetic signature coordinate functions, which is based on a linear filter defined in the frequency domain.

The parametrical model proposed in the present contribution is based on the high degree of similarity existing among the trajectory signals of real signatures in the frequency domain. In Fig. 2 some examples of DFTs of the x and y signals are shown, where we can observe that the energy of the coefficients rapidly decreases in the initial harmonics and remains constant and practically negligible from that point (marked with a vertical dashed line in Fig. 2) to the end.

This common structure of the spectrum of x and y, allows us to determine a model defined by the next parameters: i) sequence length N, ii) number of relevant spectral coefficients (i.e., in Fig. 2 number of coefficients before the dashed line), iii) magnitude of the relevant spectral coefficients, iv) magnitude of the last spectral coefficients (i.e., in Fig. 2 those after the dashed line).

In our model, the length distribution of the synthetic signatures follows that of the MCYT database[18] (comprising over 8,000 signatures), while for the rest of parameters uniform distributions between a maximum and a minimum value (extracted from the MCYT database) are assumed.

In order to generate a synthetic signature, the DFT of each of the trajectory signals is generated colouring white noise with the described parametrical model. This approach implies two simplifications: i) that all Fourier coefficients are independent, and ii) that both coordinate functions x and y are independent.

Once the synthetic DFT of both trajectory signals has been generated, we compute the Inverse DFT (IDFT) in order to obtain the coordinate functions x and y in the time domain.

2.2 The Pressure Function

The two main features defining the pressure function of a signature are: i) the number of penups (i.e., zero pressure segments of the signature) that occurred during the signing process, and ii) the placing of those penups.

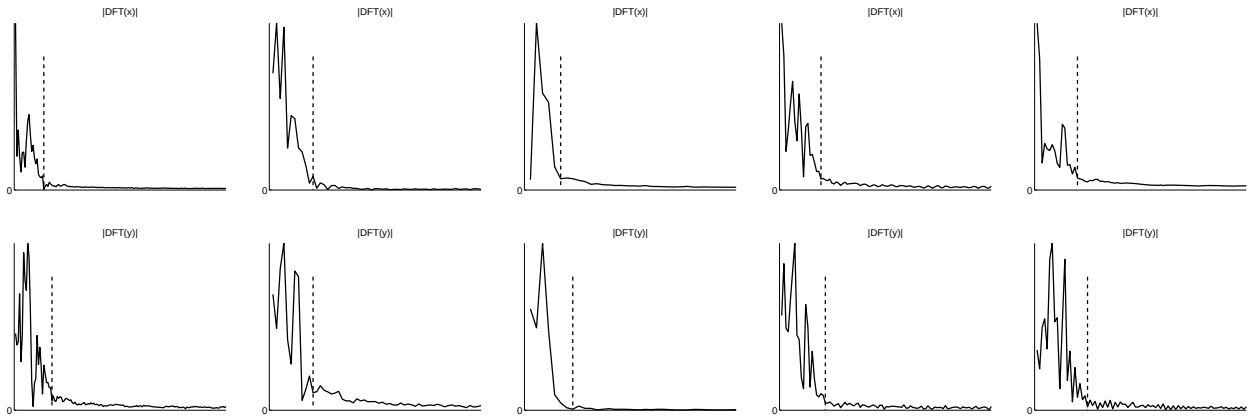


Figure 2. DFT amplitude examples of the trajectory functions x (top) and y (bottom), of 5 real signatures (from left to right).

The distribution of the number of penups was extracted from the MCYT database, and applied to the synthetic signatures according to their length (i.e., a longer signature presents a higher probability of having a large number of penups).

From an heuristical analysis of the y and p signals of real signatures we can conclude that most penups occur close to a singular point (maximum or minimum) of the y function.

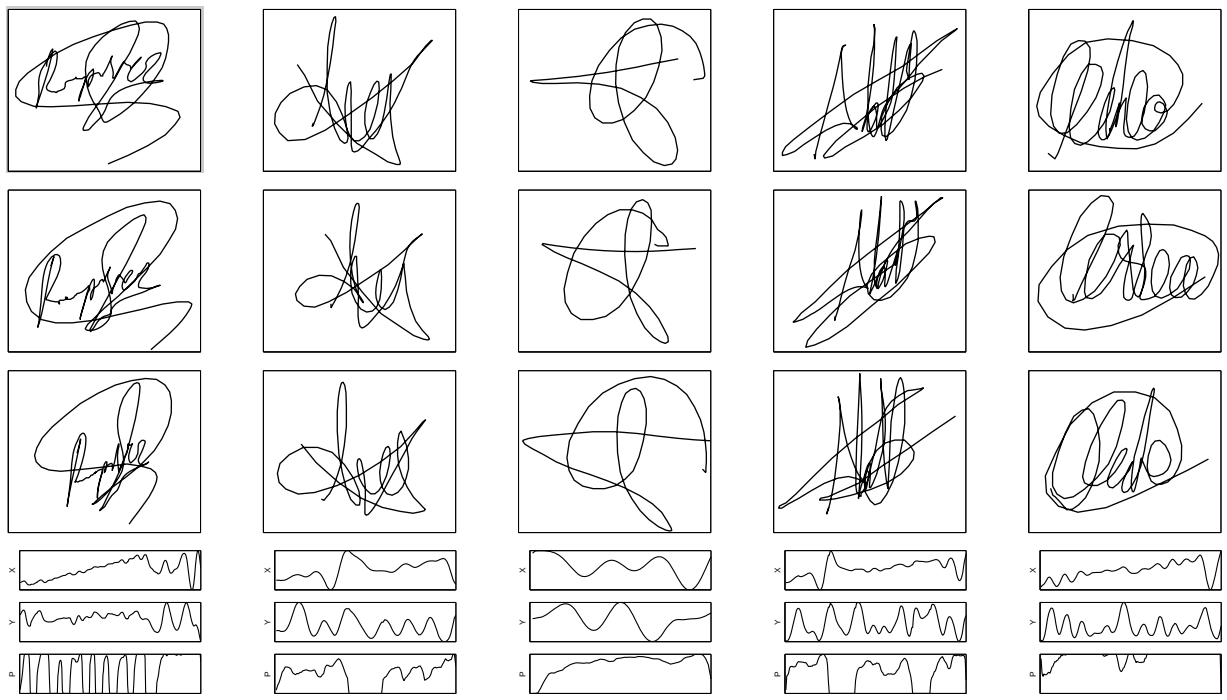
With these two premises, the penups are located through the pressure function and some maximum points (between penups) are determined randomly. In a successive step all these singular points (penups and maxima) are joined using a cubic spline interpolation algorithm. Once this initial p waveform is generated, it is processed in order to avoid undesired effects:

- Many online signature acquisition devices consider 1024 integer pressure levels, so each point of the synthetic p function is rounded to the nearest integer value, and those which exceed 1024 are set to this maximum value. The same way, those points lower than 0 are set to the penup value.
- A signature pressure signal cannot start or end with a penup. If this is the case the function is artificially changed so that the starting and ending points are non-zero elements.
- Due to the biomechanical properties of the human writing movements, penups cannot be shorter than a certain number of points (around 15 for a 100 Hz sampling rate). The pressure function is accordingly modified in order to avoid unrealistic penups.

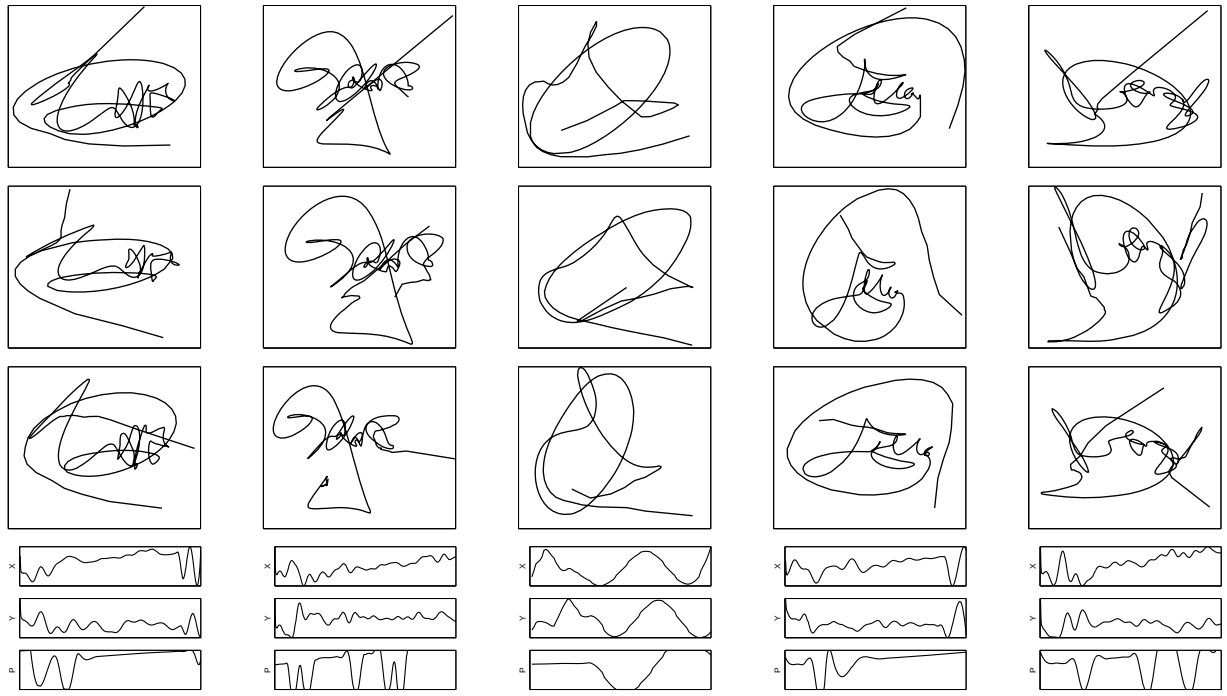
2.3 Signature Refinement in the Time Domain

Several actions are undertaken at this point to give the signature a more realistic appearance:

- Both trajectory functions are smoothed using a 10-point moving average in order to avoid possible high frequency noise.
- The x function of most left-to-right written signatures presents a general growing tendency fluctuating around a straight of fixed slope (see x function of the first real signature in Fig. 3). This behaviour is artificially produced in this step of the algorithm.
- In many cases, real signatures present a big fluctuation of their values at the end of the x and y signals, which in most cases can be identified with a round-like flourish (see x and y functions of the first real signature in Fig. 3). This final waveform is also artificially added to some signatures in this part of the algorithm.
- Additionally, translation, rotation and scaling transformations can be applied at this point if considered necessary.



Real signatures extracted from MCYT.[18]



Synthetic signatures produced with the described model-based generation algorithm.

Figure 3. Examples of real (top) and synthetic (bottom) signatures. Three samples of 5 different real and synthetic signers are shown together with the time sequences $x[n]$, $y[n]$, and $p[n]$ corresponding to the first sample.

3. MULTIPLE SAMPLES GENERATION

Once a synthetic signature (defined by its x, y, and p functions) has been created, multiple samples of that master signature are generated applying the next deformations:

- Horizontal and vertical affine scaling of all three signals.
- Duration expansion or contraction (the same length increase or decrease is applied to all three signals). The expansion/contraction factor depends on whether we want to produce intrasession ($\pm 7\%$ variation in the MCYT database), or intersession samples ($\pm 12\%$ variation in MCYT).
- Noise addition. Smoothed white noise is added to the trajectory functions. The level of noise added to generate intersession samples is about 20% higher than that inserted in the intrasession sample generation (following MCYT).

In Fig. 3 three samples of five real (top) and synthetic (bottom) signers are shown. The trajectory and pressure signals of the first sample appear below. We can observe that, although no recognizable characters can be distinguished in the synthetic signatures, their aspect and that of their time functions is quite similar to the real signatures appearance.

4. EXPERIMENTS

In addition to the observable similarity between the real and synthetic signatures appearance (patent in Fig. 3), two other experiments have been carried out in order to assess the suitability of the proposed synthetic signature generation algorithm. For that purpose a database (following the MCYT structure) of 330 synthetic signers and 25 samples per signer was generated (from now on the SSiGGeDB). The first 5 samples of each signature were generated according to the intrasession variability present in real signatures, while the remaining 20 present a higher variance in order to imitate samples acquired in different sessions.

4.1 Experiment 1: Global Features Comparison

As a first approximation to evaluate the goodness of the synthetic generation algorithm, we studied to what extent the synthetic signatures in SSiGGeDB are similar to the real signatures in MCYT according to a set of discriminative features. For that purpose, the comprehensive set of 100 global features described in[20] was extracted from each signature in MCYT and in SSiGGeDB, which comprises many of the features of the most popular works on feature-based signature verification.[8] From that 100-feature set we selected the best performing 20-parameter subset in a signature verification task (using the SFFS feature selection algorithm).[19] The resulting individual distributions of real and synthetic signatures are shown in Fig. 4, where we can observe the clear similarity between them, being in some cases (parameters 1, 21, 26, 34, and 57) practically identical.

From this result we can conclude that the most discriminant features (for verification purposes) that characterize the signature trait, are present in a very similar manner both in the real and synthetic signatures generated according to the proposed algorithm.

4.2 Experiment 2: Evaluation on a Recognition System

The performance of the synthetic signatures has been also evaluated using an HMM-based signature recognition system.[21] A 12-state and 4-mixture HMM configuration was used, with no user-dependent or score-dependent normalization of the scores. For both performance evaluations (using real and synthetic signatures) four different scenarios were considered in order to see if the behaviour of the synthetic signatures is comparable to those of the real samples, and thus can be used in the evaluation of signature verification systems:

- Number of training signatures: the performance of the system was evaluated using either 5 or 20 training signatures to compute the model of each user.
- Pressure information: the system was evaluated with and without considering the pressure information of the signatures.

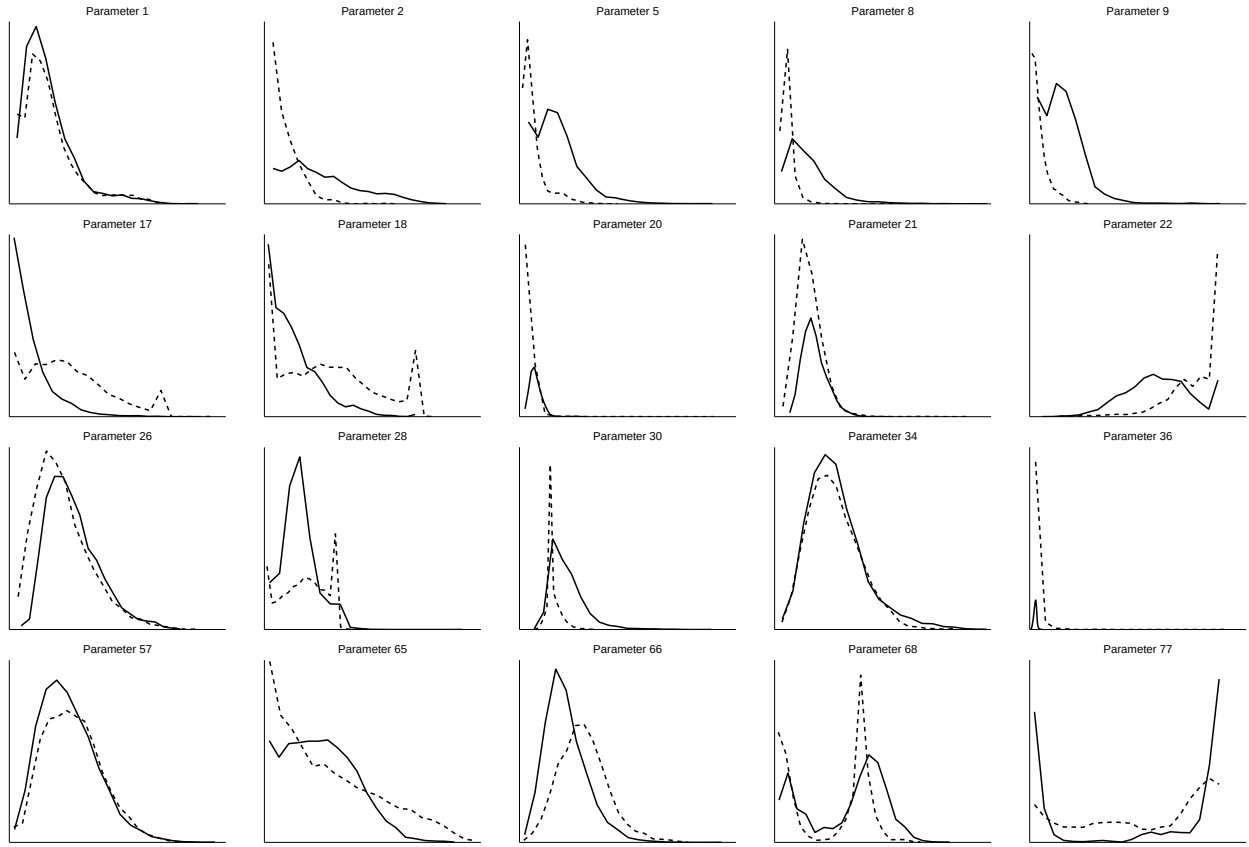


Figure 4. Histograms of real (solid lines) and synthetic (dashed lines) signatures, corresponding to the best performing 20-parameter set found by Galbally *et al.*[19] for signature verification. The parameter numeration followed by Fierrez *et al.*[20] is used, where a complete set of 100 parameters from which the best 20 were selected was introduced and discussed.

In all cases the data corpus was divided into training and test sets, where the training set comprises either 5 or 20 signatures and the test set consists of the remaining samples (this way either 330×20 or 330×5 genuine scores are produced). In order to compute the impostor scores, the trained model of each user is compared with one signature (chosen randomly) of the remaining users, thus resulting in 330×329 impostor scores.

The genuine and impostor sets of scores were computed both for real (MCYT) and synthetic signatures (SSiGGeDB), for the different scenarios considered: with and without taking into account the pressure information, and for 5 and 20 training signatures. The score distributions for all these sets of scores are shown in Fig. 5, where we can observe that, specially for the scenarios with 5 training signatures, the genuine score distribution of synthetic signatures (solid thin line) presents a bigger dispersion than that of the real signatures (solid thick line).

With those sets of scores, the EER of the system was computed and the results are shown in Table 1. Several observations can be made: *i*) the system performance on real signatures is better than with synthetic individuals, representing the latter ones a reasonable upper bound of the real performance, *ii*) in both cases (real and synthetic) there is a similar decrease in the EER when the number of training signatures increases from 5 to 20, and *iii*) for both type of signatures the inclusion of the pressure information improves the EER in a similar way.

From the two reported validation experiments we can infer that the discriminative information present in the synthetic signatures and in the real signatures, does not vary significantly. This fact makes the presented algorithm suitable to be used for the performance evaluation of automatic signature verification systems.

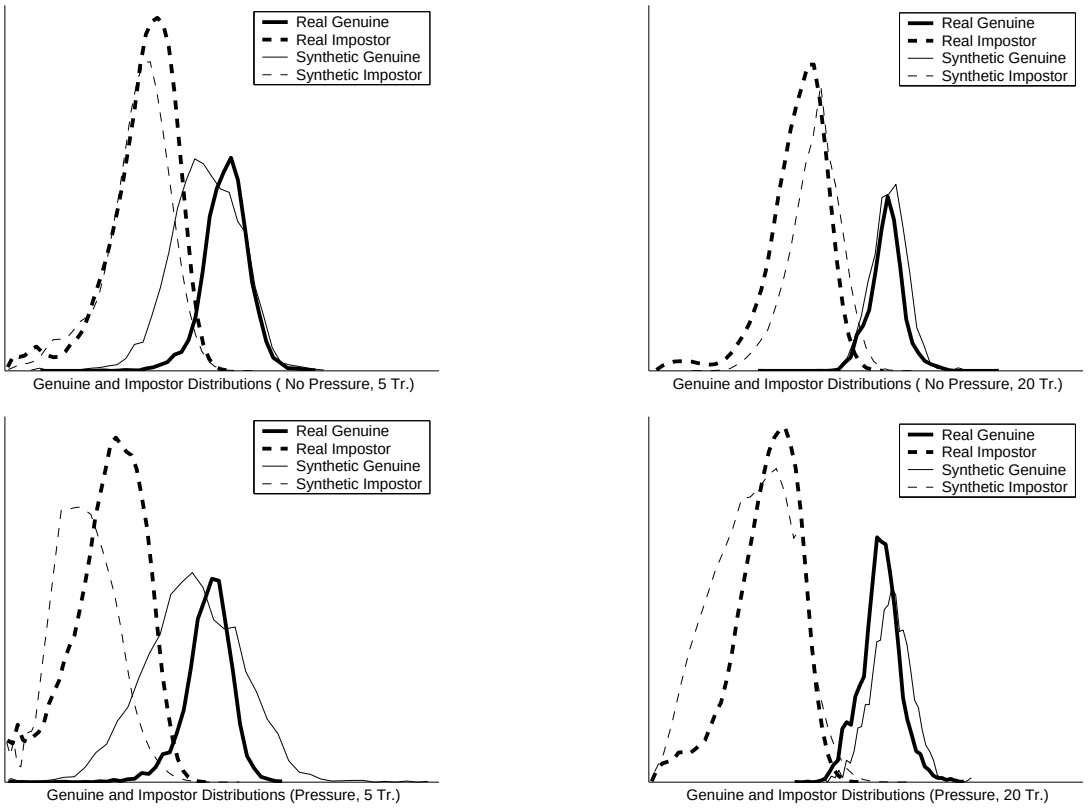


Figure 5. Score distributions of real and synthetic signatures for the different scenarios considered: with and without taking into account the pressure information and for 5 and 20 training signatures.

	EER (%)			
	No Pressure		Pressure	
	5 Tr.	20 Tr.	5 Tr.	20 Tr.
Real	4.63	1.24	3.74	0.47
Synthetic	10.41	4.17	5.83	2.03

Table 1. Performance comparison on an HMM-based signature verification system on real and synthetic signatures.

5. CONCLUSIONS

A new algorithm to generate synthetic handwritten signatures based on the spectral analysis of the signature trajectory functions has been presented. The algorithm presents the clear advantage over previously reported methods of not needing any real samples to generate new synthetic individuals. The reported experimental results show that the synthetically generated signatures, in addition to presenting a very realistic appearance, possess very similar characteristics to those that enable the recognition of real signatures. The proposed algorithm can be used as an efficient tool for the evaluation of automatic signature verification systems, as it can rapidly and easily generate large amounts of realistic data.

In addition to evaluation and vulnerability assessment tasks, the proposed synthetic generation method can also be useful as a development tool in other biometric applications where the data scarcity is a key issue. In particular it can be used to:

- Generate data from multiple signers for signature recognition approaches using data-driven machine learn-

ing, where large amounts of data to train the classifier are needed.

- Generate multiple samples of given users in order to overcome the shortage of data in verification and identification training.
- Study in depth the nature, properties and limitations of the signature signal in order to identify individuals (e.g., individuality studies), to increase the robustness of the current recognition systems, or to obtain more robust signatures against forgeries.

ACKNOWLEDGMENTS

J. G. is supported by a FPU Fellowship from Spanish MEC. J. F. postdoctoral research is supported by a Marie Curie Fellowship from the European Commission. This work was supported by Spanish MEC under project TEC2006-13141-C03-03.

REFERENCES

[1] Dutoit, T., [*An introduction to text-to-speech synthesis*], Kluwer Academic Publishers (2001).
[2] Cappelli, R., [*Handbook of Fingerprint Recognition*], ch. Synthetic Fingerprint Generation, 203–231, Springer (2003).
[3] Zuo, J., Schmid, N. A., et al., "On generation and analysis of synthetic iris images," *IEEE Trans. IFS* **2**, 77–90 (2007).
[4] Lin, A. and Wang, L., "Style-preserving english handwriting synthesis," *Pattern Recognition* **40**, 2097–2109 (2007).
[5] Popel, D. V., [*Synthesis and analysis in biometrics*], ch. Signature analysis, verification and synthesis in pervasive environments, 31–63, World Scientific (2007).
[6] Cappelli, R., Maio, D., et al., "Performance evaluation of fingerprint verification systems," *IEEE Trans. on PAMI* **28**(1), 3–18 (2006).
[7] Galbally, J., Fierrez, J., et al., "On the vulnerability of fingerprint verification systems to fake fingerprint attacks," in [*Proc. IEEE ICCST*], 130–136 (2006).
[8] Fierrez, J. and Ortega-Garcia, J., [*Handbook of biometrics*], ch. On-line signature verification, 189–209, Springer (2008).
[9] Oliveira, C., Kaestner, C. A., et al., "Generation of signatures by deformations," in [*Proc. BSDIA*], 283–298 (1997).
[10] Djioua, M., O'Reilly, C., and Plamondon, R., "An interactive trajectory synthesizer to study outlier patterns in handwriting recognition and signature verification," in [*Proc. ICPR*], (2006).
[11] Rabasse, C., Guest, R. M., and Fairhurst, M. C., "A method for the synthesis of dynamic biometric signature data," in [*Proc. ICDAR*], (2007).
[12] Richiardi, J., "Skilled and synthetic forgeries in signature verification: large-scale experiments," tech. rep., Institute of Electrical Engineering, Swiss Federal Institute of Technology, Lausanne (2008).
[13] Munich, M. E. and Perona, P., "Visual identification by signature tracking," *Trans. PAMI* **25**, 200–217 (2003).
[14] Ballard, L., Lopresti, D., and Monrose, F., "Forgery quality and its implications for behavioral biometric security," *IEEE Trans. on Systems, Man, and Cybernetics* **37**, 1107–1118 (2007).
[15] Bezine, H., Kefi, M., and Alimi, M., "On the beta-elliptic model for the control of the human arm movement," *International Journal of Pattern Recognition* **21**, 5–19 (2007).
[16] Djioua, M., Plamondon, R., et al., "Deterministic and evolutionary extraction of delta-lognormal parameters: performance comparison," *International Journal of Pattern Recognition* **21**, 21–41 (2007).
[17] Kholmatov, A. and Yanikoglu, B., "An individuality model for online signatures using global Fourier descriptors," in [*Proc. BTHI V*], **6944**(1) (2008).
[18] Ortega-Garcia, J., Fierrez-Aguilar, J., et al., "MCYT baseline corpus: a bimodal biometric database," *IEE Proc. VISP* **150**(6), 391–401 (2003).

[19] Galbally, J., Fierrez, J., and Ortega-Garcia, J., "Performance and robustness: a trade-off in dynamic signature verification," in [*Proc. ICASSP*], 1697–1700 (2008).
[20] Fierrez-Aguilar, J., Nanni, L., et al., "An on-line signature verification system based on fusion of local and global information," in [*Proc. of IAPR AVBPA*], 523–532, Springer LNCS-3546 (2005).
[21] Fierrez, J., Ortega-Garcia, J., et al., "HMM-based on-line signature verification: feature extraction and signature modeling," *Pattern Recognition Letters* **8**, 2325–2334 (2007).