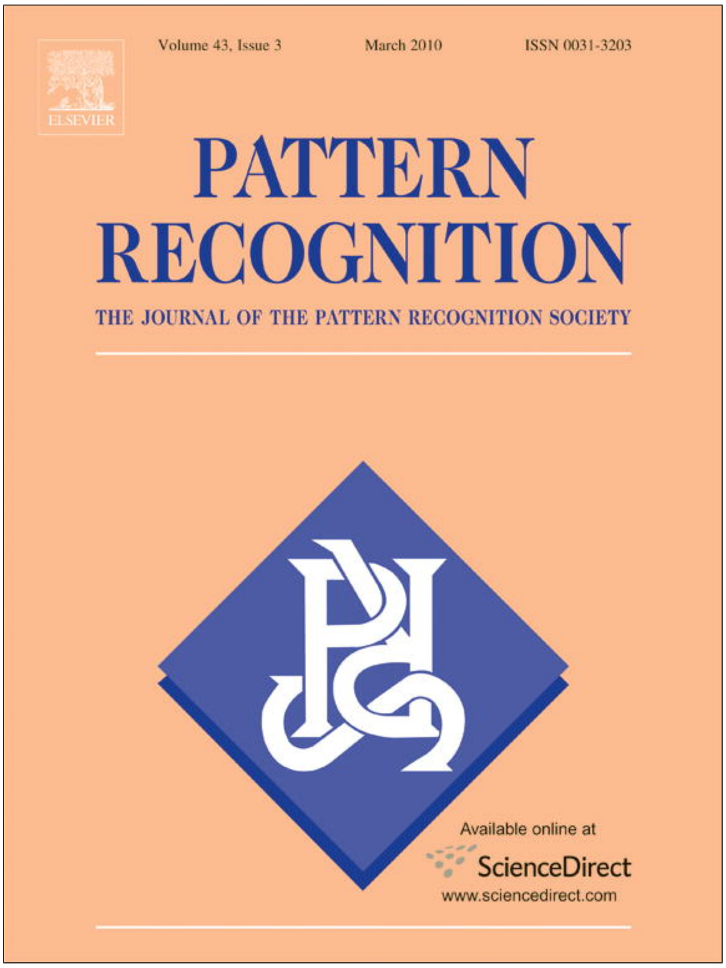

Volume 43, Issue 3
March 2010
ISSN 0031-3203
ELSEVIER
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
Available online at
ScienceDirect
www.sciencedirect.com

Pattern Recognition 43 (2010) 1027–1038

Contents lists available at ScienceDirect

Pattern Recognition

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

On the vulnerability of face verification systems to hill-climbing attacks

Javier Galbally [a,*], Chris McCool [b], Julian Fierrez [a], Sebastien Marcel [b], Javier Ortega-Garcia [a]

[a] Biometric Recognition Group - ATVS, EPS, Universidad Autonoma de Madrid, C/ Francisco Tomas y Valiente, 11-28049 Madrid, Spain
[b] IDIAP Research Institute, Rue Marconi 19-1920 Martigny, Switzerland

ARTICLE INFO

Article history:
Received 5 May 2009
Received in revised form
21 July 2009
Accepted 28 August 2009

Keywords:
Biometrics
Face recognition
Security
Vulnerabilities
Hill-climbing

ABSTRACT

In this paper, we use a hill-climbing attack algorithm based on Bayesian adaption to test the vulnerability of two face recognition systems to indirect attacks. The attacking technique uses the scores provided by the matcher to adapt a global distribution computed from an independent set of users, to the local specificities of the client being attacked. The proposed attack is evaluated on an eigenface-based and a parts-based face verification system using the XM2VTS database. Experimental results demonstrate that the hill-climbing algorithm is very efficient and is able to bypass over 85% of the attacked accounts (for both face recognition systems). The security flaws of the analyzed systems are pointed out and possible countermeasures to avoid them are also proposed.

1. Introduction

Automatic access of persons to services is becoming increasingly important in the information era. This has resulted in the establishment of a new research and technology area known as biometric recognition, or simply biometrics [1]. The basic aim of biometrics is to discriminate automatically between subjects—in a reliable way and according to some target application—based on one or more signals derived from physical or behavioral traits, such as fingerprint, face, iris, voice, hand, or written signature.

Biometric technology presents several advantages over classical security methods that are based on a pass-phrase (Personal Identification Number or password) or on a physical key (or access card) [2,3]. A major disadvantage of traditional authentication systems is that they cannot discriminate between impostors who have illegally acquired the privileges to access a system and the genuine user. Furthermore, in biometric systems there is no need for the user to remember difficult PIN codes that could be easily forgotten or to carry a key that could be lost or stolen.

Despite their advantages, biometric systems are still vulnerable to external attacks which could decrease their level of security. Thus, it is of utmost importance to analyze the vulnerabilities of biometric systems, in order to find their limitations and to develop useful countermeasures for foreseeable attacks. Furthermore, the vulnerability study carried out in the present work can be of great use for other parties working in the biometric field such as developers or security evaluators. In particular, the interest for the analysis of security vulnerabilities has surpassed the scientific community and different standardization initiatives at international level have emerged in order to deal with the problem of security evaluation in biometric systems, such as the common criteria (CC) through different supporting documents [15], or the biometric evaluation methodology (BEM) [16]. The present research work can be of great help to further develop these ongoing security evaluation standardization efforts, and for independent institutions in charge of objectively asserting the level of security offered to the final user by face recognition systems.

In [4] Ratha identified and classified eight possible attack points for biometric recognition systems. These vulnerability points, depicted in Fig. 1, can be broadly divided into two groups:

- *Direct attacks*: In [4] the possibility to generate synthetic biometric samples (for instance, speech, fingerprints or face images) in order to illegally access a system was discussed, and defined as the first vulnerability point in a biometric security system. These attacks at the sensor level are referred to as *direct attacks* and require no specific knowledge about the system (e.g., matching algorithm, feature extraction process or feature vector format). Furthermore, the attack is carried out in the analog domain, outside the digital limits of the system, so digital protection mechanisms (digital signature or watermarking) cannot be used. Some previous works have studied the robustness of biometric systems to direct attacks, specifically finger- and iris-based systems [5–7].

[*] Corresponding author.
E-mail addresses: javier.galbally@uam.es (J. Galbally), christopher.mccool@idiap.ch (C. McCool), julian.fierrez@uam.es (J. Fierrez), sebastien.marcel@idiap.ch (S. Marcel), javier.ortega@uam.es (J. Ortega-Garcia).

0031-3203/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2009.08.022

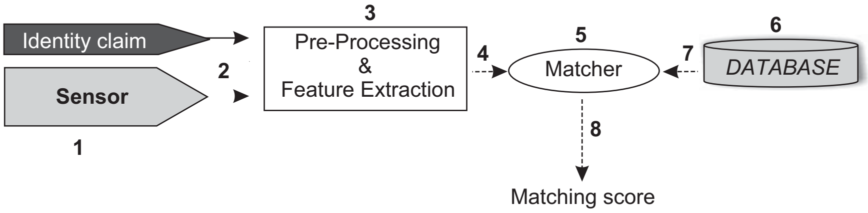


Fig. 1. Architecture of an automated biometric verification system. Possible attack points are numbered from 1 to 8.

- *Indirect attacks*: This group includes all the remaining seven points of attack identified in Fig. 1. Attacks 3 and 5 might be carried out using a Trojan Horse that bypasses the feature extractor and the matcher, respectively. In attack 6 the system database is manipulated (a template is changed, added or deleted) in order to gain access to the application. The remaining points of attack (2, 4, 7 and 8) are thought to exploit possible weak points in the communication channels of the system by extracting, adding or changing information from them. In this case the intruder needs to have some additional information about the internal working of the recognition system and, in most cases, physical access to some of the application components (feature extractor, matcher or database) is required.

Some efforts have been made to study the robustness of biometric systems against indirect attacks. In [8] a model-based attack which is capable of reconstructing the user's face images from the matching scores is presented. The method has the strong constraint of needing a large number of real face images to initialize the algorithm.

Apart from [8], most of the works studying the vulnerability of biometric systems to indirect attacks use some type of variant of the hill-climbing algorithm presented in [9]. In that preliminary work, a basic hill climbing attack was tested over a simple image recognition system using filter-based correlation. This attack takes advantage of the score given by the matcher to iteratively change a synthetically created template until the score exceeds a fixed decision threshold and thereby gain access to the system.

Two hill-climbing attacks to a standard and Match-on-Card minutiae-based fingerprint verification systems have been reported in [11,12], respectively. In these attacks a synthetic random minutiae template is presented to the input of the matcher and, according to the score generated, the random template is iteratively changed until the system returns a positive verification. The minutiae in the template are modified one at a time and the change is only stored if the score returned by the matcher improves the previous one, otherwise it is discarded. The changes included in the modification scheme are adding, substituting, changing or deleting a minutia, which make the attack not applicable to any other biometric system different from a minutiae-based fingerprint recognition system.

Adler proposed a hill-climbing attack to a face recognition system in [10]. The input image, which is selected from an arbitrary set of real face images, is modified using an independent set of eigenfaces (which makes it applicable only to face recognition systems) until the desired matching score is attained. This work reported results on a PCA-based face recognition system and showed that after 3000 iterations, a score corresponding to a very high similarity confidence (99.8%) was reached. The success rate of the attack (how many accounts were broken out of the total attacked) or the operating point of the system is not given, so the results are difficult to interpret or compare.

Most of the hill-climbing approaches are all highly dependent on the technology used, only being usable for a very specific type of matcher. However, in [13] a hill-climbing algorithm based on Bayesian adaptation, which can be applied to attack different biometric systems, was presented and tested using an on-line signature verification system. In the present contribution this attack is successfully applied to two automatic face recognition systems thus showing its big attacking potential and its ability to adapt to different biometric systems and matchers which use fixed length feature vectors of real numbers and delivering real similarity (or dissimilarity) scores.

Two case studies are presented in this work where several aspects of the attack are investigated. The first one examines the effectiveness of the technique on an eigenface-based verification system while the second uses a more advanced Gaussian mixture model (GMM) parts-based approach. For both case studies the experiments are conducted on the XM2VTS database [14] and it is shown that the attack is able to bypass over 85% of the accounts attacked for the best configuration of the algorithm found. Furthermore, the hill-climbing approach is shown to be faster than a brute-force attack for all the operating points evaluated, as well as being capable of reconstructing the user's face image from the similarity scores, without using any real face images to initialize the algorithm. As a result, the proposed algorithm has vulnerability implications related to both security and privacy issues of the users.

The paper is structured as follows. The hill-climbing attack algorithm used in the experiments is described in Section 2, while the two attacked systems are presented in Section 3. The database and experimental protocol followed are described in Section 4. The results on the eigenface-based system and the GMM system are detailed in Sections 5.1 and 5.2, respectively. Conclusions are finally drawn in Section 6.

2. Bayesian hill-climbing algorithm

Problem statement. Consider the problem of finding a K-dimensional vector $\mathbf{y}^*$ which, compared to an unknown template $\mathcal{C}$ (in our case related to a specific client), produces a similarity score bigger than a certain threshold δ, according to some matching function J, i.e., $J(\mathcal{C},\mathbf{y}^*)>\delta$. The template can be another K-dimensional vector or a generative model of K-dimensional vectors.

Assumptions. Let us assume:

- That there exists a statistical model G (K-variate Gaussian with mean $\boldsymbol{\mu}_G$ and diagonal covariance matrix $\boldsymbol{\Sigma}_G$, with $\boldsymbol{\sigma}_G^2=\mathrm{diag}(\boldsymbol{\Sigma}_G)$), in our case related to a background set of users, overlapping to some extent with $\mathcal{C}$.
- That we have access to the evaluation of the matching function $J(\mathcal{C},\mathbf{y})$ for several trials of $\mathbf{y}$.

Algorithm. The problem of finding $\mathbf{y}^*$ can be solved by adapting the global distribution G to the local specificities of template $\mathcal{C}$, through the following iterative strategy:

1. Take N samples ($\mathbf{y}_i$) of the global distribution G, and compute the similarity scores $J(\mathcal{C},\mathbf{y}_i)$, with $i=1,\ldots,N$.

2. Select the M points (with $M<N$) which have generated highest scores.
3. Compute the local distribution $L(\boldsymbol{\mu}_L,\boldsymbol{\sigma}_L)$, also K-variate Gaussian, based on the M selected points.
4. Compute an adapted distribution $A(\boldsymbol{\mu}_A,\boldsymbol{\sigma}_A)$, also K-variate Gaussian, which trades off the general knowledge provided by $G(\boldsymbol{\mu}_G,\boldsymbol{\sigma}_G)$ and the local information given by $L(\boldsymbol{\mu}_L,\boldsymbol{\sigma}_L)$. This is achieved by adapting the sufficient statistics as follows:

$$\boldsymbol{\mu}_A=\alpha\boldsymbol{\mu}_L+(1-\alpha)\boldsymbol{\mu}_G \tag{1}$$

$$\boldsymbol{\sigma}_A^2=\alpha(\boldsymbol{\sigma}_L^2+\boldsymbol{\mu}_L^2)+(1-\alpha)(\boldsymbol{\sigma}_G^2+\boldsymbol{\mu}_G^2)-\boldsymbol{\mu}_A^2 \tag{2}$$

5. Redefine $G=A$ and return to step 1.

In Eqs. (1) and (2), $\boldsymbol{\mu}^2$ is defined as $\boldsymbol{\mu}^2=\mathrm{diag}(\boldsymbol{\mu}\boldsymbol{\mu}^T)$, and α is an adaptation coefficient in the range [0,1]. The algorithm finishes either when one of the N similarity scores computed in step 2 exceeds the given threshold δ or when the maximum number of iterations is reached.

In the above algorithm there are two key concepts not to be confused, namely: (i) number of *iterations* (n_{it}), which refers to the number of times that the statistical distribution G is adapted and (ii) number of *comparisons* (n_{comp}), which denotes the total number of matchings carried out through the algorithm. Both numbers are related through the parameter N, being $n_{comp}=N\cdot n_{it}$.

3. Face verification systems attacked

The described Bayesian hill-climbing algorithm is used to test the robustness against this type of attacks of two different face verification systems, one based on eigenfaces [17], and a second using GMM with a part-based representation of the face [18]:

- *Eigenface-based system*: The face verification system used for the evaluation of the hill-climbing attack is based on the well known eigenfaces technique introduced by Turk and Pentland in [17]. This algorithm applies eigen-decomposition to the covariance matrix of a set of M vectorized training images $\overline{\mathbf{x}}_i$. In statistical pattern recognition this technique is referred to as PCA [19]. This method has become a *de facto* standard for face verification and was used to present initial results for the recent face recognition grand challenge evaluation [20].
  The first similarity measure used to compare PCA based features was the Euclidean distance, however, several other similarity measures have been later proposed and studied [21]. The evaluated system uses cropped face images of size 64 × 80 to train a PCA vector space where 80% of the variance is retained. This leads to a system where the original image space of 5120 dimensions is reduced to 91 dimensions ($K=91$). Similarity scores are then computed in this PCA vector space using the standard correlation metric, $d(\mathbf{x},\mathbf{y})=1-[(\mathbf{x}-\mu_{\mathbf{x}})\cdot(\mathbf{y}-\mu_{\mathbf{y}})]/\sigma_{\mathbf{x}}\sigma_{\mathbf{y}}$, as it showed the best performance out of the tested similarity measures.
- *GMM parts-based system*: The GMM parts-based system used in the evaluation tesselates the 64 × 80 images into 8 × 8 blocks with a horizontal and vertical overlap of 4 pixels. This tessalation process results in 285 blocks and from each block a feature vector is obtained by applying the discrete cosine transform (DCT); from the possible 64 DCT coefficients only the first 15 coefficients are retained ($K=15$). The blocks are used to derive a world GMM Ω_w and a client GMM Ω_c [18]. Experimentation found that using a 512 mixture component GMM gave optimal results.
  When performing a query, or match, the average score of the 285 blocks from the input image is used. The DCT feature vector from each block v_i (where $i=1\ldots285$) is matched to both Ω_w and Ω_c to produce a log-likelihood score. These scores are then combined using the log-likelihood ratio, $S_{llr,j}=\log[P(v_j|\Omega_c)]-\log[P(v_j|\Omega_w)]$, and the average of these scores is used as the final score, $S_{GMM}=\frac{1}{285}\sum_{j=1}^{285}S_{llr,j}$. This means that the query template can be considered to be a feature matrix formed by 285 fifteen dimensional vectors (representing each of the blocks in the image).

4. Database and experimental protocol

4.1. The XM2VTS database

The experiments are carried out on the XM2VTS face database [14], comprising 295 users. The database was acquired in four time-spaced capture sessions in which two different face images of each client were taken under controlled conditions (pose and illumination) to complete the total 295 × 8 = 2360 samples of the database. Two evaluation protocols are defined for this database, the Lausanne Protocol 1 and 2 (LP1 and LP2). In Fig. 2 some examples of images that can be found in the XM2VTS are shown.

4.2. Performance evaluation

The performance of the evaluated systems is computed based on the LP2 protocol. This protocol is chosen as the training and evaluation data are drawn from independent capture sessions.

According to LP2 the database is divided into: (i) a training set comprising the samples of the two first sessions of 200 clients (used to compute the PCA transformation matrix, and the world GMM Ω_w, respectively) and (ii) a test set formed by the fourth session images of the previous 200 users (used to compute the client scores), and all the eight images of 70 different users with which the impostor scores are calculated. As a result of using the same subjects for PCA training and client enrollment, the system performance is optimistically biased, and therefore harder to attack than in a practical situation (in which the enrolled clients may not have been used for PCA training). This means that the results presented in this paper are a conservative estimate of the attack's success rate. In Fig. 4 a general diagram showing the LP2 evaluation protocol is given (although defined by LP2, the development set was not used in our experiments).

In the case of the eigenface-based system, the final score given by the system is the average of the p scores obtained after matching the input vector to the p templates of the attacked client model $\mathcal{C}$, while in the GMM system the p templates are used to estimate the parameters of the client GMM (Ω_c). In Fig. 3 we can see the system false acceptance rate (FAR) and false rejection rate (FRR) curves for the eigenface-based system (left) and for the GMM system (right), using the described protocol with $p=4$ enrollment templates. The eigenface-based system presents an equal error rate (EER) of 4.71%, while the GMM system shows a better performance with a 1.24% EER. The three operating points where the hill-climbing algorithm is evaluated (corresponding to FAR = 0.1%, FAR = 0.05%, and FAR = 0.01%) are also highlighted. These operating points correspond to a low, medium, and high security application according to [22].

Fig. 2. Examples of the images that can be found in XM2VTS.

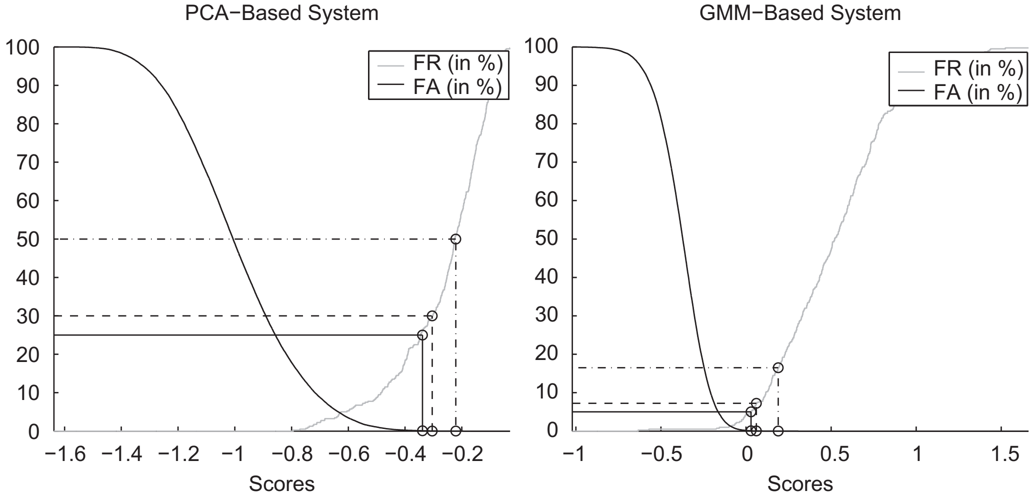
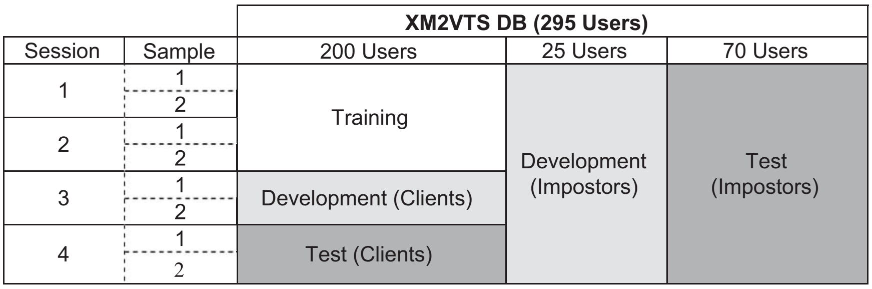


Fig. 3. FAR and FRR curves for the eigenface-based system (left) and the GMM-based system (right).



		XM2VTS DB (295 Users)		
Session	Sample	200 Users	25 Users	70 Users
1	1	Training	Development (Impostors)	Test (Impostors)
	2			
2	1			
	2			
3	1	Development (Clients)		
	2			
4	1	Test (Clients)		
	2			

Fig. 4. Diagram showing the partitioning of the XM2VTS database according to the LP2 protocol (which was used in the performance evaluation of the present work).

4.3. *Experimental protocol for the attacks*

In order to generate the user accounts to be attacked using the hill-climbing algorithm, we used the train set defined by LP2 (i.e., samples corresponding to the first two sessions of 200 users).

The initial K-variate distribution G of the algorithm was estimated using part or all the samples (depending on the experiment) from the impostors in the test set (70 users) defined in LP2 (referred to in the rest of the work as generation set). This way, there is no overlap between the attacked set of users (200 accounts), and the subjects used to initialize the algorithm, which could lead to biased results on the success rate (SR) of the attack. The SR is defined as the number of accounts broken A_b by the attack (i.e., accounts where the hill-climbing scheme reaches the decision threshold δ), divided by the total number of accounts attacked $A_T = 200$. Thus, $SR = A_b/A_T$. In Fig. 5 the partitioning of the database used for the attacks is shown.

5. Experiments

The goal of these experiments is to study the vulnerability of automatic face recognition systems to hill-climbing attacks. This is achieved by examining the effectiveness of the Bayesian-based hill-climbing algorithm in attacking two different face recognition systems at several operating points. By performing these attacks it will also be studied the ability of the Bayesian-based hill-climbing algorithm to adapt, not only to different matchers, but also to other biometric traits (it was already shown to be successful attacking an on-line signature verification system in [13]).

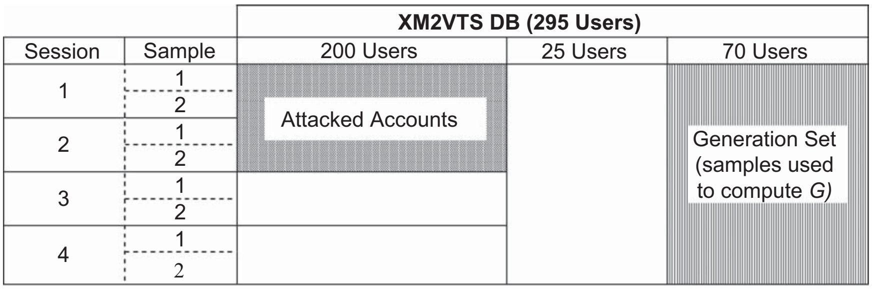


Fig. 5. Diagram showing the partitioning of the XM2VTS database followed in the attacks protocol.

Two case studies are presented for the attacks on the two separate face verification systems. The first case study examines the effectiveness of the Bayesian-based hill-climbing attack on the eigenface-based system (Section 5.1). The second study uses the previously found optimal configuration to attack the GMM parts-based system (Section 5.2).

5.1. Case study 1: attacking an eigenface-based face verification system

In the first set of experiments, we study the effect of varying the three parameters of the algorithm (N, M, and α) on the success rate (SR) of the attack over the eigenface-based system (described in Section 3). The objective is to reach an optimal configuration where the number of broken accounts is maximized, while minimizing the average number of comparisons (n_{comp}) needed to reach the fixed threshold δ. As described in Section 2, the above mentioned parameters denote: N the number of sampled points of the adapted distribution at a given iteration, M the number of top ranked samples used at each iteration to adapt the global distribution, and α is an adaptation coefficient which varies from $[0\ldots1]$.

The importance of the initial distribution G is also examined by evaluating the attack performance when a smaller number of real samples is used to compute G. The case where G is randomly selected is also examined.

When presenting results the brute-force approach is used to provide a baseline to compare with the hill-climbing algorithm. We compare n_{comp} with the number of matchings necessary for a successful brute-force attack at the operating point under consideration ($n_{bf} = 1/\text{FAR}$). However, it should be noticed that the proposed hill-climbing algorithm and a brute-force attack are not fully comparable as the latter requires much greater resources (e.g., a database of thousands of samples).

5.1.1. Analysis of N and M (sampled and retained points)

For the initial evaluation of the algorithm an operating point of FAR $= 0.01\%$ was fixed (this FAR leads to an FRR of 50%). This FAR implies that an eventual brute-force attack would be successful, on average, after 10,000 comparisons. Given this threshold the algorithm was executed for different values of N and M (fixing $\alpha = 0.5$) and the results are given in Table 1. The maximum number of iterations (n_{it}) allowed for the algorithm appears in brackets. This value changes according to N in order to maintain constant the maximum number of comparisons permitted ($n_{comp} = N \cdot n_{it}$). In plain text we show the success rate of the attack (in % over the total 200 accounts tested), while the average number of comparisons needed for a successful attack is represented in bold.

Table 1
Success rate (in %) of the hill-climbing attack for increasing values of N (number of sampled points) and M (best ranked points).

	N				
	10 (2500)	25 (1000)	50 (500)	100 (250)	200 (125)
M					
3	84.5 **5162**	86.0 **4413**	86.0 **4669**	86.0 **5226**	86.0 **6296**
5	81.5 **5796**	*86.0* ***4275***	86.0 **4512**	86.0 **5022**	86.0 **5988**
10		85.5 **4534**	86.0 **4540**	86.0 **5019**	86.0 **5941**
25			86.0 **5213**	86.0 **5379**	86.0 **6256**
50				86.0 **6455**	86.0 **6934**
100					86.0 **8954**

The maximum number of iterations allowed is given in brackets. The success rate (in %) appears in plain text, while the average number of iterations needed to break an account appears in bold. The best configuration of parameters N and M is highlighted in italic and bold italic.

Examining Table 1 the optimal configuration for these parameters is $[N = 25, M = 5]$ (highlighted in gray). For this point, the number of accounts broken is maximized (86%) and n_{comp} is minimized (4275). This minimum represents less than half of the expected number of matchings required for a successful brute-force attack ($n_{bf} = 1/\text{FAR} = 10,000$).

Further analysis of the results indicates that selecting the best possible N has a deeper impact on the speed of the attack than choosing a good value for M. This is because N represents the number of scores produced at each iteration of the attack and consequently has a direct impact on the number of comparisons performed n_{comp}.

It can also be drawn from the results presented in Table 1 that choosing a value such that $N > M$ provides a better efficiency (in terms of n_{comp}) than if $M \simeq N$ (the sub-sampling of the local distribution is too general and so the speed of the attack is reduced) or $N \gg M$ (the sub-sampling of the local distribution is too specific which again reduces the speed of the attack).

Irrespective of how N and M are optimized the number of accounts broken by the attack remains stable. For almost all the configurations evaluated 86% of the accounts were broken (172 out of a total of 200). Further examining this result it was found that the 28 clients who remain robust to the attack are the same in all cases.

To search for an explanation, the 28 unbroken client models (comprising the four images of the first two database sessions) were matched to the other four images of the user (those corresponding to sessions three and four). It was found that none of the client models produced a score high enough to enter the system, which means that these 28 clients would not be suitable

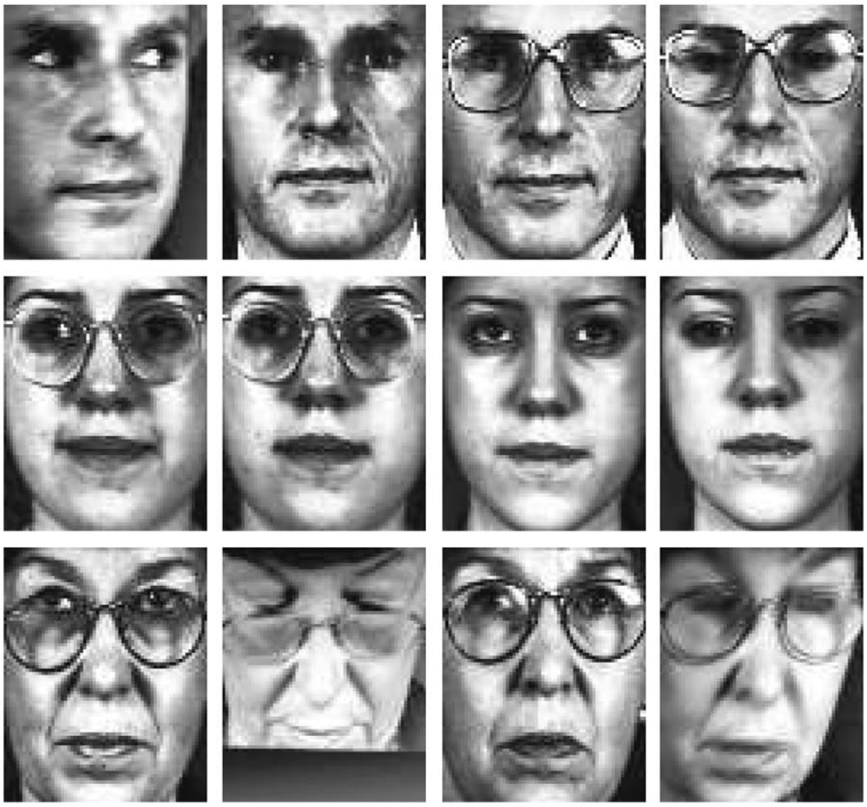


Fig. 6. The four enrollment images (columns) constituting the model of three of the unbroken accounts (rows).

Table 2
Success rate (in %) of the hill-climbing attack for increasing values of α and for $[N,M]=[25,5]$.

α	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
SR (%)	0	84.5	86.0	86.0	86.0	86.0	86.0	81.0	71.5	51.0	20.0
n_{comp}	**25,000**	**6468**	**5121**	**4617**	**4381**	**4275**	**4380**	**4990**	**7901**	**10,404**	**14,154**

The success rate (in %) appears in plain text, while the average number of iterations needed to break an account appears in bold.

for face recognition under the considered system working at the selected operating point. We can then conclude that the attack successfully broke all the models that would be used in a real application. In Fig. 6 the enrollment images which form three of the resistant accounts are shown. In all cases we can observe a great variance among the samples of a given model (glasses/not glasses, different poses, and blurred images).

5.1.2. Analysis of α (adaptation coefficient)

For the optimal configuration of N and M the effect of varying α on the performance of the attack is studied. This parameter is changed from 0 (only the global distribution G is taken into account) to 1 (only the local distribution L affects the adaptation stage). The results are presented in Table 2 where the success rate of the attack appears in plain text (%), while the average number of comparisons needed for a successful attack is shown in bold.

From Table 2 it can be seen that the optimal point is $\alpha=0.5$ (where both the number of accounts broken is maximized and the number of comparisons needed minimized). This corresponds to the case where both the global and local distributions are given approximately the same importance. As in the previous experiment, it can be noticed that 14% of the accounts (the same 28 clients as in the previous experiments) is never bypassed as a consequence of the large user intra-variability.

5.1.3. Analysis of the initial distribution G

In the previous experiments the K-variate initial distribution G was computed using the two images from the first session of the 70 users comprised in the generation set (see Fig. 5). In this section the effect of estimating G using different number of samples, and a random initialization of G, are both explored.

In Table 3 we show how the performance of the attack varies depending on the number of samples used to estimate this distribution G, for the best configuration of the attack $[N,M,\alpha]=[25,5,0.5]$. As the generation set comprises 70 users, for numbers of images smaller than 70, one sample per subject (randomly selected from the generation set) was used, while for 70 images or larger numbers, 1, 2, 4, and 8 samples from each subject are used. In all cases, the resulting multivariate Gaussian G results in $[-0.8<\mu_i<0.5]$ and $[0.2<\sigma_i<18]$, where μ_i and σ_i are, respectively, the mean and variance of the i-th dimension, with $i=1\ldots91$.

No real samples are used in the random initialization, where G corresponds to a multivariate Gaussian of zero mean and variance one.

Table 3
Success rate (in %) of the hill-climbing attack for increasing number of samples used to compute the initial distribution G.

Number of real samples used to compute G							Random ($\mu=0, \sigma=1$)
5	10	35	70	140	280	560	
86.0 **4353**	86.0 **4307**	86.0 **4287**	86.0 **4283**	86.0 **4279**	86.0 **4285**	86.0 **4281**	86.0 **4492**

N, M, and α are set to 25, 5, and 0.5, respectively. The success rate (in %) appears in plain text, while the average number of iterations needed to break an account appears in bold.

Table 4
Results of the attack for different points of operation and the best configuration found of the attacking algorithm ($N=25$, $M=5$, $\alpha=0.5$).

	Operating points (in %)		
	FAR = 0.1, FRR = 25	FAR = 0.05, FRR = 30	FAR = 0.01, FRR = 50
SR (in %)	99.0	98.5	86.0
n_{comp}	**840**	**1068**	**4492**
n_{bf}	1000	2000	10,000

The success rate is given in plain text (over a total of 200 accounts), and n_{comp} in bold. The average number of matchings needed for a successful brute-force attack (n_{bf}) is also given for reference.

From the results shown in Table 3 we can see that the number of samples used to compute the initial distribution G has little effect on the performance of the attack. In fact, the experiment shows that the algorithm can be successfully run starting from a general initial distribution G of zero mean and unit variance. This means that an attacker would not need to have any real face images to carry out the attack (on the studied system), which is in stark contrast to a brute force attack which requires a large database to perform a successful attack.

5.1.4. Analysis of different operating points

Using the best configuration $[N, M, \alpha] = [25, 5, 0.5]$ and starting from a general initial distribution G of zero mean and unit variance, the algorithm was evaluated in two additional operating points of the system (see Fig. 3). The two additional operating points are: (i) FAR $= 0.05\%$, which implies $n_{bf} = 2000$ and leads to FRR $= 30\%$, and (ii) FAR $= 0.1\%$, which implies $n_{bf} = 1000$ and leads to FRR $= 25\%$. Results are given in Table 4.

Smaller values of the FAR imply a bigger value of the threshold δ to be reached by the algorithm, which causes a rise in the average number of iterations required for a successful attack. However, the results in Table 4 demonstrate that this technique is effective across multiple operating points. In all cases the number of comparisons needed to break the system (using the Bayesian hill-climbing attack) is lower than that of a brute force attack. The hill-climbing approach has the added advantage that it does not need any real face images to initialize the attack.

5.1.5. Graphical analysis of the attack

In order to illustrate graphically how the hill-climbing algorithm works we repeated the attack for the best configuration $[N, M, \alpha] = [25, 5, 0.5]$ at a high security operating point (FAR $= 0.01\%$). To visualize the hill climbing attack we present the results using the Euclidean distance as the similarity measure. This metric provides very similar results to those obtained with the standard correlation metric (in terms of the SR of the attack and n_{comp}), however, due to the different characteristics of both measures (the standard correlation is angle based) the Euclidean distance provides a more intuitive visual insight into the effect of the hill-climbing attack, as can be observed in Figs. 7 and 8.

In Figs. 7 and 8 two examples of broken and non-broken accounts (corresponding to two of the users presented in Fig. 6) are shown. For each of the examples the evolution of the score through the iterations of the algorithm is depicted. Highlighted in each example are six points, including the first and the last ones, of the iterative process marked with letters A through to F. The dashed line represents the objective value to be reached (i.e., the threshold δ). The two upper faces correspond to one of the original images of the attacked user and the reconstructed image of a K-dimensional eigenface template (where part of the information has been lost because of the dimensionality reduction). The sequence of the six faces below corresponds to the feature vectors that produced each of the six scores marked with A through to F. The first point A is produced by randomly sampling the estimated general distribution $\mathcal{G}$ and the last point F represents the image which is able to break the system. These two figures show that the algorithm can be used not only as a break-in strategy but also as a method to accurately reconstruct the client's face image (with the privacy issues that this entails).

In Figs. 7 and 8 we can observe that the hill-climbing algorithm starts from a totally random face which is iteratively modified to make it resemble as much as possible to the PCA projection of the attacked users face labeled as "Original-PCA" (this effect cannot be observed as clear when using the standard correlation metric). In both cases (broken and non-broken accounts) the attack successfully finds a final image which is very similar to the objective face, however, in the case of the accounts resistant to the attack, the threshold is not reached as a consequence of the large user intra-variability, which leads to low scores even when compared with images of the same client.

5.2. Case study 2: attacking a GMM face verification system

In order to attack the GMM-based system, the best configuration of the algorithm found in the previous experiments was used ($N=25$, $M=5$, and $\alpha=0.5$). The default operating point to attack the system corresponds to FAR $= 0.01\%$ (this means that a brute force attack would need on average to be successful $n_{bf} = 10,000$ matchings), which leads to FRR $= 16\%$.

Two different approaches to the problem of attacking the GMM system are tested in these experiments:

- *Single block search*: This attack searches for one block to break the client's account. As explained in Section 3, the client score Sc is computed by taking the average score from all the blocks, therefore, if we are able to find one good matching block and replicate it for all the other blocks we should be able to produce a score high enough to be granted access. With these premises, this attack uses the Bayesian adaptation to search for one 15 dimensional vector which is repeated 285 times in order to produce the final synthetic template capable of breaking the system.
- *Multiple block search*: In this case we search for a unique set of vectors which are capable of breaking into the client's account. Like the single block search this attack undertakes a search in a 15 dimensional space, however, in this case 285 random vectors (of 15 dimensions) are sampled to generate the synthetic client template. As before, when performing the Bayesian adaptation the average of the M best synthetic templates is used to produce the vectors $\boldsymbol{\mu}_L$ and $\boldsymbol{\sigma}_L$. The fact that we are looking for a greater number of vectors than in the

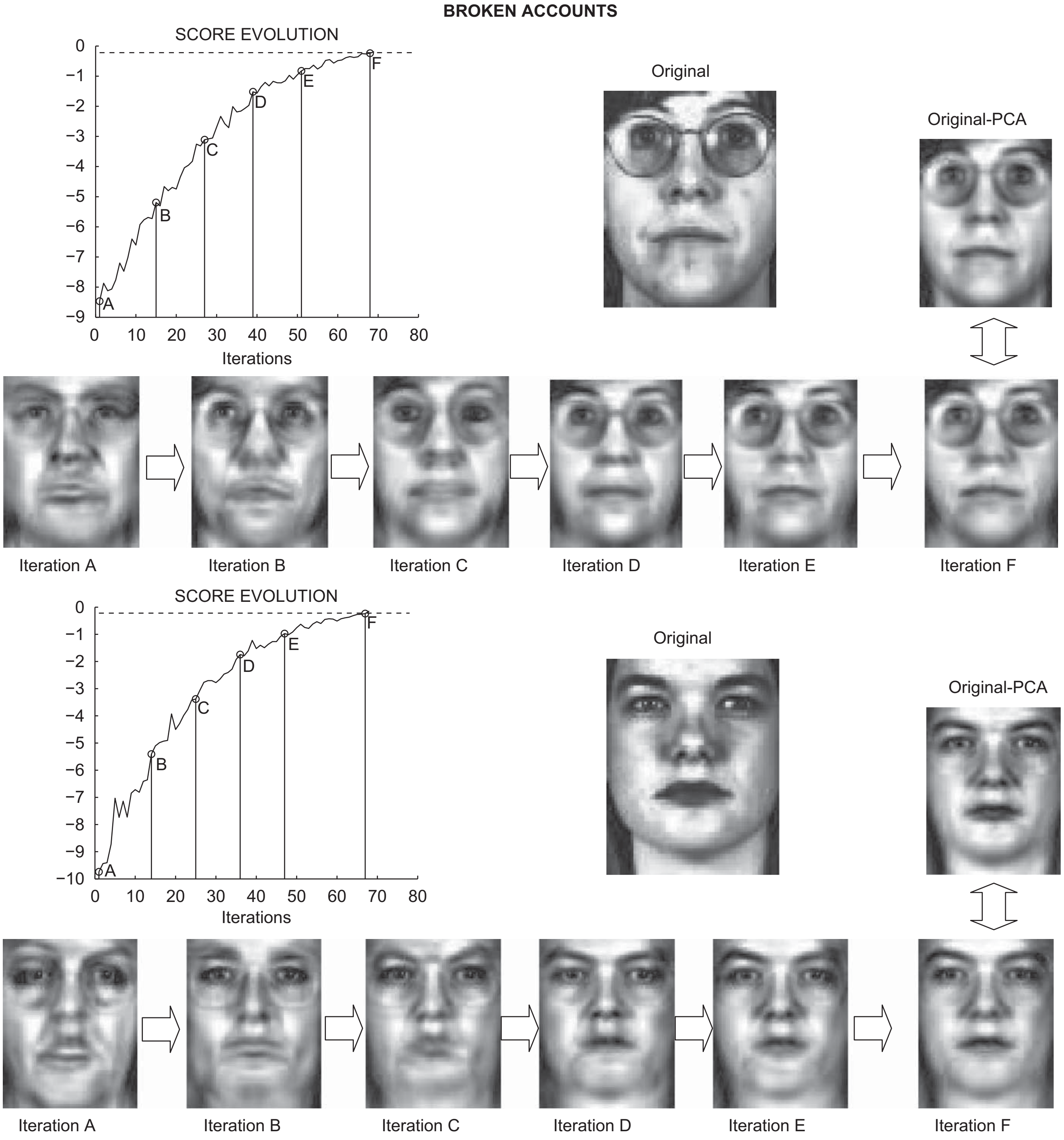


Fig. 7. Examples of the evolution of the score and the synthetic eigenfaces through the iterations of the attack for broken and accounts. The dashed line represents the objective threshold.

single block search makes the multiple block search more difficult to accomplish and also more difficult to detect.

5.2.1. *Experiments starting from an average initial distribution G*

For these experiments we computed an initial distribution G representing the average block (i.e., mean and average of the 15 dimensional blocks found in several images). The distribution was computed using a different number of images selected from the generation set defined in the attack protocol (see Fig. 5). For numbers of images smaller than 70, one sample per user (randomly selected) is picked, while for larger numbers (140, 280, and 560) 2, 4, and 8 samples per subject are selected, respectively. In Tables 5 and 6 the results for the single and multiple block search approaches are shown.

For the single block search all the accounts are broken at the first iteration of the attack (at each iteration 25 comparisons are computed). This means that the Bayesian adaptation hill-climbing algorithm is not necessary and that the system can be broken using synthetic templates built replicating 285 times a random

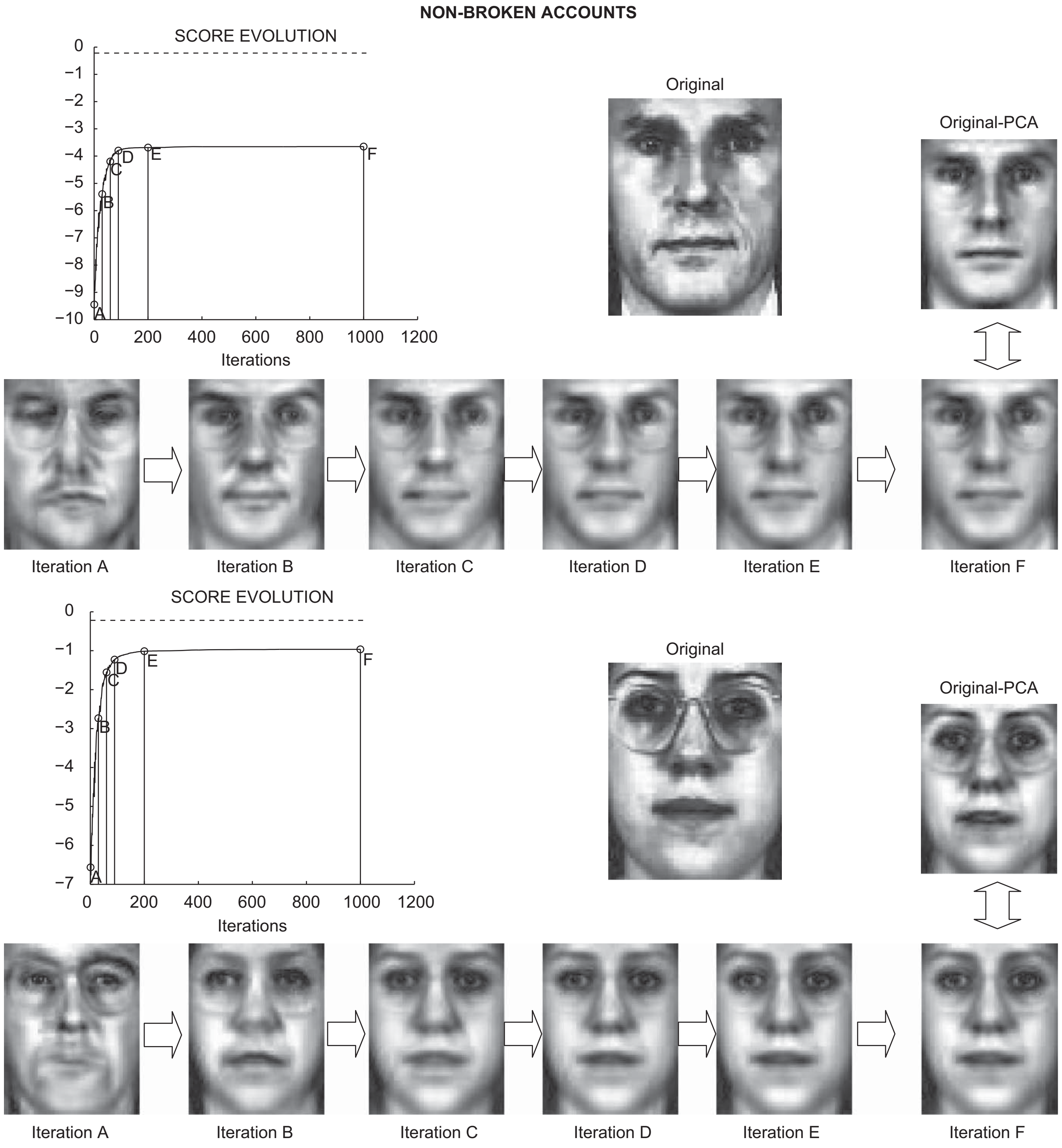


Fig. 8. Examples of the evolution of the score and the synthetic eigenfaces through the iterations of the attack for non-broken and accounts. The dashed line represents the objective threshold.

average block estimated using as few as five images. This serious security flaw can be countermeasured by checking if all the blocks in the template trying to access the system are different.

The multiple block search attack has almost a 100% success rate regardless of the number of images used to compute the initial distribution G. However, for this attack we would need, on average, around 1200 comparisons (corresponding to 55 iterations of the attack) to break the system. This represents less than one-sixth of the matchings required by a successful brute force attack ($n_{bf}10,000$) with the added advantage that just five real face images are needed to perform the hill-climbing attack. Although the multiple block search is slower than the single block search approach, in this case countermeasuring the attack is significantly more difficult as all the vectors, which form the synthetic template, are different amongst themselves.

5.2.2. *Experiments starting from a random initial distribution G*

The GMM-based system was also attacked starting from a random initial distribution G with zero mean and unit variance. For the single block search approach 98% of the accounts (out of the total 200 tested) were bypassed, and the average number of

Table 5
Success rate (in %) of the hill-climbing attack under single (top) and multiple (bottom) block search, for increasing number of real samples used to compute the initial distribution G.

	Number of real samples used to compute G						
	5	10	35	70	140	280	560
Sing. block search	100	100	100	100	100	100	100
	25	**25**	**25**	**25**	**25**	**25**	**25**
Mult. block search	100	100	100	100	99.5	100	100
	1031	**1025**	**1631**	**1514**	**1328**	**1293**	**1254**

The success rate (in %) appears in plain text, while the average number of iterations needed to break an account appears in bold.

Table 6
Results of the attack for different points of operation and the best configuration found of the attacking algorithm ($N=25$, $M=5$, $\alpha=0.5$).

	Operating points (in %)		

	FAR = 0.1,
	FRR = 5
		FAR = 0.05,
		FRR = 7
			FAR = 0.01,
			FRR = 16

Sing. block search	100	100	98
	123	**413**	**1102**
Mult. block search	100	100	100
	724	**1835**	**3016**
n_{bf}	1000	2000	10,000

The success rate is given in plain text (over a total 200), and n_{comp} in (bold). The average number of matchings needed for a successful brute-force attack (n_{bf}) is also given for reference.

matchings needed to enter the system was 1102. Although that success rate is very high, we can observe in Fig. 9 that the hill-climbing is not working properly as the score remains unaltered and equal to zero throughout the iterations (there is no increasing or *hill-climbing* effect) until at one point it very rapidly (two or three iterations) reaches the objective value (shown with a dashed line).

This behavior can be explained by the fact that the score given by the system is the substraction of the client and the world scores (see Section 3). As the synthetic templates are built duplicating a block randomly selected from a general distribution G, their appearance is completely different to that of a face and so both similarity scores (those obtained from the world and client model) are the same, leading to a zero final score. As the final score obtained by all the synthetic templates is the same (zero), we have no feedback as about the local distribution L (representing those templates which are more similar to the attacked one). Therefore, the algorithm ends up doing a random search until at some point one of the templates produces (by chance) a non-zero score.

Even though this attack is the equivalent of a random search it successfully breaks the system at the first attempt (corresponding to 25 matchings) for 43% of the tested accounts. Therefore, this security breach should be taken into account when designing countermeasures (e.g., checking that all the blocks of the template are different) for final applications.

The above experiments were repeated using the multiple block search scheme. In this case, all 200 accounts were bypassed and the average number of comparisons needed to break the system was 3016. In Fig. 10 it can be observed that the hill-climbing algorithm is able to produce the desired increasing effect in the

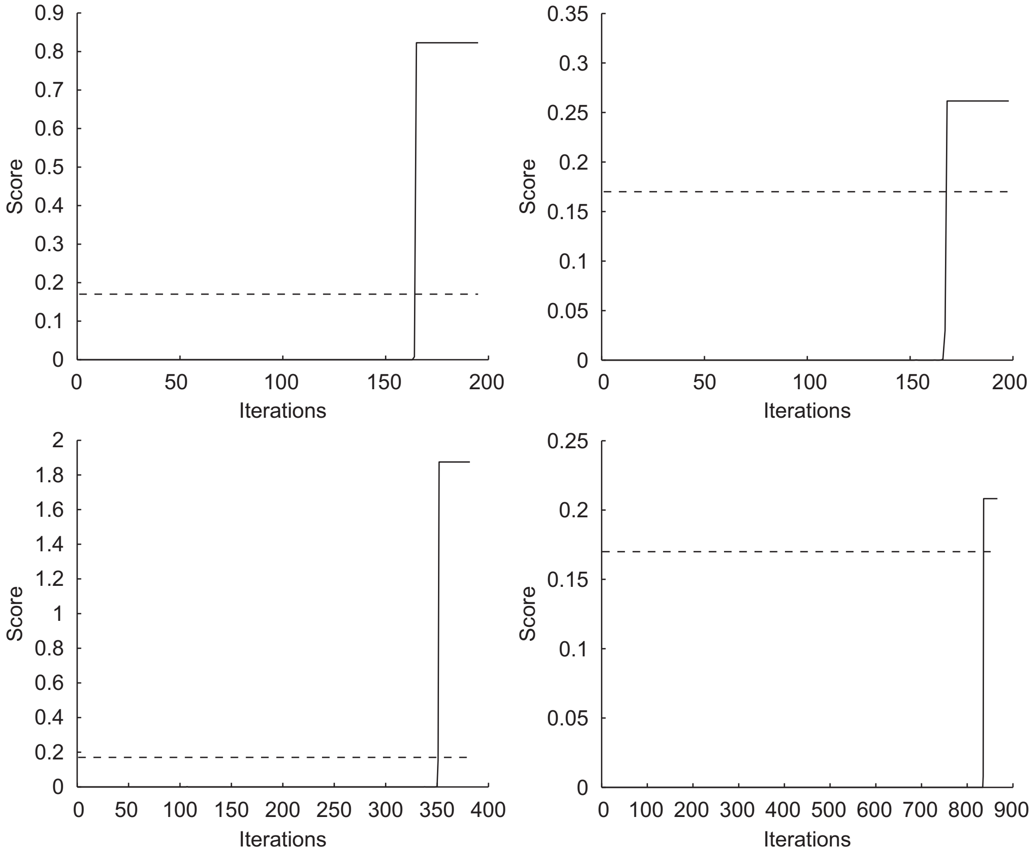


Fig. 9. Evolution of the score for four of the broken accounts using the single block search approach on the GMM-based face verification system. The dashed line represents the objective threshold.

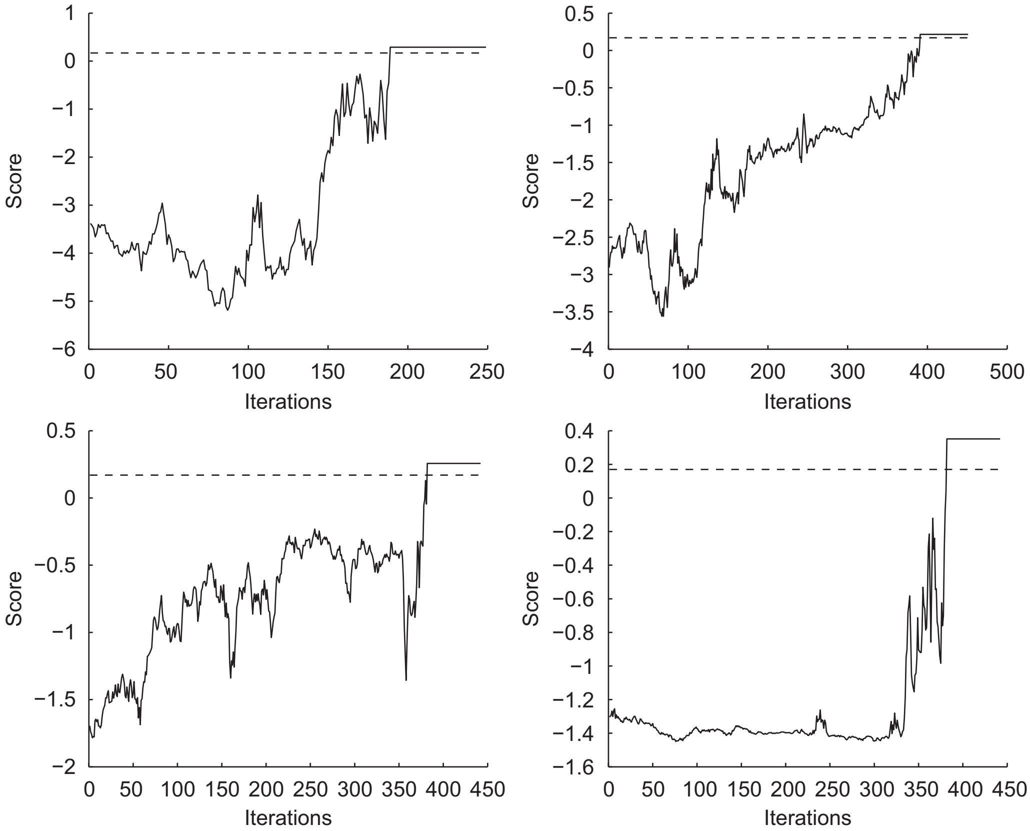


Fig. 10. Evolution of the score for four of the broken accounts using the multiple block search approach on the GMM-based face verification system. The dashed line represents the objective threshold.

score throughout the iterations. We can see that the synthetic templates produce a negative final score (they get a better matching score from the world model than from the client model, $S = Sc - Sw$) and thus, the algorithm gets the necessary feedback to iteratively improve the estimate of the vector distribution G. Again, this approach is slower than the single block search, but on the other hand it is more difficult to countermeasure as all the image blocks are different amongst themselves.

5.2.3. Analysis of different operating points

The GMM-based system was evaluated at two additional operating points, these being: (i) FAR $= 0.05\%$, which implies $n_{bf} = 2000$ and leads to FRR $= 7\%$, and (ii) FAR $= 0.1\%$, which implies $n_{bf} = 1000$ and leads to FRR $= 5\%$. For these experiments the initial distribution G was chosen as a Gaussian distribution with zero mean and unit variance and the two different attacking approaches (single block search and multiple block search) were tested.

The results indicate that the Bayesian hill-climbing attack is effective for all of the operating points. The number of broken accounts remains unaltered (100% for all cases) and, the same as in the PCA-based system study, the number of comparisons needed to break the system is always lower than that of a brute force attack.

6. Conclusions

The robustness of two different face verification systems (one PCA-based and one working on GMMs) against a hill-climbing attack based on Bayesian adaptation has been studied. Experimental results show that the two face verification systems studied are highly vulnerable to this type of attack, with over an 85% success rate for all of the attacks; even when no real images were used to initialize the algorithm. Furthermore, the attack showed its ability to reconstruct the user's real face image from the scores, thus arising security issues concerning the privacy of the client.

The performance of the Bayesian hill-climbing algorithm was compared to a brute force attack. It was found that the Bayesian hill-climbing attack is more efficient under all tested conditions. In addition, it is worth noting that the resources required by both approaches differ greatly. In order to perform an efficient brute-force attack, the attacker must have a database of more than a thousand real different templates, while the hill-climbing approach does not need any real templates to be successful.

It has also been found that the GMM-based system, although its overall performance is significantly better than the PCA-based system, is very vulnerable to random attacks carried out with templates formed by a replicated random or average block. This important security flaw can be solved by incorporating to the systems a mechanism to detect duplicated patterns in the image.

At the same time, the present study points out the serious risk that the Bayesian-based hill-climbing algorithm represents as it has been successfully applied not only to different matchers but also to different biometric traits (in [13] it was shown to be an effective method to attack an on-line signature verification system). Thus, this threat should be studied when designing biometric security systems working with fixed length feature vectors of real numbers and delivering real similarity scores.

Applying this technique to a multi-class classifier and not a verification system (two-class problem) is not straight forward. Therefore, applying this technique directly to a multi-class SVM or probabilistic neural network represents a challenging attacking scenario that will be the source of future research.

Acknowledgments

J.G. is supported by a FPU Fellowship from Spanish MEC and J.F. is supported by a Marie Curie Fellowship from the European Commission (EC). J.G. would also like to thank the IDIAP Research Institute for hosting him during the development of the present research work.

This work has been supported by Spanish MEC under Project TEC2006-13141-C03-03 and the European NoE Biosecure.

This work has been performed under the MOBIO project 7th Framework Research Programme of the European Union (EU), Grant agreement number: 214324. The authors would like to thank the EU for the financial support and the partners within the consortium for a fruitful collaboration. For more information about the MOBIO consortium visit http://www.mobioproject.org.

References

[1] A.K. Jain, A. Ross, S. Pankanti, Biometrics: a tool for information security, IEEE Transactions on Information Forensics and Security 1 (2006) 125–143.
[2] A.K. Jain, P. Flynn, A. Ross (Eds.), Handbook of Biometrics, Springer, Berlin, 2008.
[3] J. Wayman, A. Jain, et al., Biometric Systems. Technology, Design and Performance Evaluation, Springer, Berlin, 2005.
[4] N. Ratha, J. Connell, R. Bolle, An analysis of minutiae matching strength, in: Proceedings of the Audio- and Video-Based Biometric Person Authentication (AVBPA), 2001, pp. 223–228.
[5] T. van der Putte, J. Keuning, Biometrical fingerprint recognition: don't get your fingers burned, in: Proceedings of the IFIP Conference on Smart Card Research and Advanced Applications (CARDIS), 2000, pp. 289–303.
[6] J. Galbally, J. Fierrez, et al., On the vulnerability of fingerprint verification systems to fake fingerprint attacks, in: Proceedings of the IEEE of International Carnahan Conference on Security Technology (ICCST), 2006, pp. 130–136.
[7] A. Pacut, A. Czajka, Aliveness detection for iris biometrics, in: Proceedings of the IEEE of International Carnahan Conference on Security Technology (ICCST), 2006, pp. 122–129.
[8] P. Mohanty, S. Sarkar, R. Kasturi, From scores to face templates: a model-based approach, Pattern Analysis and Machine Intelligence 29 (2007) 2065–2078.
[9] C. Soutar, Biometric system security. 〈http://www.bioscrypt.com/assets/security_soutar.pdf〉.
[10] A. Adler, Sample images can be independently restored from face recognition templates, in: Proceedings of the Canadian Conference Electrical and Computing Engineering (CCECE), vol. 2, 2003, pp. 1163–1166.
[11] U. Uludag, A.K. Jain, Attacks on biometric systems: a case study in fingerprints, in: Proceedings of the SPIE-IE, vol. 5306, 2004, pp. 622–633.
[12] M. Martinez-Diaz, J. Fierrez, et al., Hill-climbing and brute force attacks on biometric systems: a case study in match-on-card fingerprint verification, in: Proceedings of the IEEE of International Carnahan Conference on Security Technology (ICCST), 2006, pp. 151–159.
[13] J. Galbally, J. Fierrez, J. Ortega-Garcia, Bayesian hill-climbing attack and its application to signature verification, Proceedings of the IAPR International Conference on Biometrics ICB, Lecture Notes in Computer Science, vol. 4642, Springer, Berlin, 2007, pp. 386–395.
[14] K. Messer, J. Matas, et al., XM2VTSDB: the extended M2VTS database, in: Proceedings of the IAPR Audio- and Video-Based Biometric Person Authentication (AVBPA), 1999.
[15] CC: Common Criteria for Information Technology Security Evaluation, V3.1., 2006.
[16] BEM: Biometric Evaluation Methodology, V1.0, 2002.
[17] M.A. Turk, A.P. Pentland, Face recognition using eigenfaces, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1991, pp. 586–591.
[18] F. Cardinaux, C. Sanderson, S. Marcel, Comparison of MLP and GMM classifiers for face verification on xm2vts, in: Proceedings of the IAPR International Conference on Audio- and Video-based Biometric Person Authentication (AVBPA), 2003.
[19] K. Fukunaga, Introduction to Statistical Pattern Recognition, Academic Press, New York, 1990.
[20] J. Phillips, P. Flynn, et al., Overview of the face recognition grand challenge, in: Proceedings of the International Conference on Computer Vision and Pattern Recognition (ICCVPR), 2005.
[21] W.S. Yambor, B.A. Draper, J.R. Beveridge, Analyzing PCA-based face recognition algorithms: eigenvector selection and distance measures, in: Proceedings of the WEECV, 2000.
[22] ANSI X9.84-2001, Biometric Information Management and Security, 2001.

About the Author—JAVIER GALBALLY received the M.Sc. and degree in electrical engineering in 2005 from Universidad de Cantabria, Spain. In 2005 he joined the Biometrics Recognition Group - ATVS at the Universidad Autonoma de Madrid, where he is currently working as an assistant researcher pursuing the Ph.D. degree. His research interests include security and biometric systems evaluation based both on fingerprint and signature. He is currently involved in National and European projects focused on biometrics such as Biosecur-ID and the European NoE Biosecure.

About the Author—CHRIS is a post-doctoral researcher at the Idiap Research Institute. He received his PhD in 2007 from Queensland University of Technology in Australia. His research interests include pattern recognition and computer vision with a particular empahsis on biometrics, 3D face verification, 2D face verification and face detection.

About the Author—JULIAN FIERREZ received the M.Sc. and the Ph.D. degrees in telecommunications 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 currently holds a Marie Curie Postdoctoral Fellowship. As part of that fellowship, he has recently spent a 2-year period as visiting researcher at Michigan State University in USA. His research interests and areas of expertise include signal and image processing, pattern recognition, and biometrics, with emphasis on signature and fingerprint verification, multi-biometrics, biometric databases, and system security.

For more information visit: http://arantxa.ii.uam.es/∼jfierrez/index.html.

About the Author—SÉBASTIEN MARCEL is a senior research scientist at the Idiap Research Institute. He manages research projects and supervises a research team in the field of biometric person recognition. He has obtained his Ph.D. in signal processing from "Universite de Rennes I" in France (2000). Prior to joining IDIAP, he was a research and development engineer at France Telecom R&D now Orange Labs. Dr. Sébastien Marcel is currently interested in multiple aspects of biometric person recognition (face detection and recognition, speaker verification, EEG-based authentication), but also in man–machine interaction (hand gesture recognition) and content-based multimedia indexing and retrieval. For more information see: http://www.idiap.ch/∼marcel.

About the Author—JAVIER ORTEGA-GARCIA received the M.Sc. degree in electrical engineering (Ingeniero de Telecomunicacion) in 1989; and the Ph.D. 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 ATVS research group. He is currently a professor at the Escuela Politecnica Superior, Universidad Autnoma de Madrid, where he teaches Digital Signal Processing and Speech Processing courses. He also hold a Ph.D. degree course in Biometric Signal Processing. His research interests are focused on biometrics signal processing: speaker recognition, face recognition, fingerprint recognition, on-line signature verification, data fusion and multimodality in biometrics. 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. Since 2008 he is a fellow member of the IEEE.