# Hill-Climbing Attack to an Eigenface-Based Face Verification System

Javier Galbally, Julian Fierrez, and Javier Ortega-Garcia
Biometric Recognition Group–ATVS, EPS, UAM
C/ Francisco Tomas y Valiente 11, 28049 Madrid, Spain
Email: {javier.galbally, julian.fierrez, javier.ortega}@uam.es

Chris McCool, and Sebastien Marcel
IDIAP Research Institute, Rue Marconi 19
1920 Martigny, Switzerland
Email: {chris.mccool, sebastien.marcel}@idiap.ch

***Abstract*—We use a general hill-climbing attack algorithm based on Bayesian adaption to test the vulnerability of an Eigenface-based approach for face recognition against indirect attacks. The attacking technique uses the scores provided by the matcher to adapt a global distribution, computed from a development set of users, to the local specificities of the client being attacked. The proposed attack is evaluated on an Eigenface-based verification system using the XM2VTS database. The results show a very high efficiency of the hill-climbing algorithm, which successfully bypassed the system for over 85% of the attacked accounts.**

## I. Introduction

Biometric security systems present several advantages over traditional security approaches [1] and consequently they are currently being introduced in many applications, including: access control, sensitive data protection, and on-line tracking systems. However, in spite of these advantages they can be exposed to external attacks which can 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.

There are some previous works that study the robustness of biometric systems against attacks carried out against the input of the matching module. In [2] 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.

Most of the existing works studying the vulnerabilities of biometric systems to attacks against the inner modules of the system, apart from [2], use some type of variant of the hill-climbing algorithm presented in [3]. Some examples include an attack to a face-based system in [4], and to a standard and a Match-on-Card minutiae-based fingerprint verification systems in [5] and [6] respectively. These attacks take advantage of the score given by the matcher to iteratively change a synthetically created template until the similarity score exceeds a fixed decision threshold and thereby gain access to the system.

Most of these hill-climbing approaches are all highly dependent on the technology used, only being usable for very specific type of matchers. However, in [7] a general hill-climbing algorithm based on Bayesian adaptation was presented and tested using a signature verification system.

In the present contribution this general attack is successfully applied to an automatic face recognition system based on eigenfaces thus proving its biometric independency and its ability of adapt to different matchers which use fixed length feature vectors. The experiments are conducted on the XM2VTS database [8], from which 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 without using any real face images to initialize the algorithm.

The paper is structured as follows. The hill-climbing attack algorithm used in the experiments is outlined in Sect. II, while the attacked system is presented in Sect. III. The database and experimental protocol followed are described in Sect IV. The results are detailed in Sect. V and conclusions are finally drawn in Sect. VI.

## II. Bayesian Hill-Climbing Attack Algorithm

The term hill-climbing designates an attack in which the similarity score given by the matcher is used to iteratively modify a synthetically generated template, or group of templates, until the verification threshold ($\delta$) is reached.

In the present contribution we use the Bayesian approach for hill-climbing presented in [7]. The core idea behind the algorithm is to iteratively adapt a known global distribution to the local specificities of the unknown user being attacked. For this purpose, a pool of users is used to compute the general statistical model $\mathcal{G}$, which is sampled $N$ times. Each sampled point $\mathbf{y_i}$ ($i = 1 \dots N$) in the distribution is compared with the client being attacked $\mathcal{C}$, generating $N$ similarity scores $s_i = J(\mathcal{C}, \mathbf{y}_i)$. The $M$ points which have generated higher scores are then used to compute a local distribution $\mathcal{L}$, which is used to generate an adapted distribution $\mathcal{A}$, that trades off (according to a parameter $\alpha$) the general knowledge provided by $\mathcal{G}$ and the local information given by $\mathcal{L}$. The global distribution is then redefined as $\mathcal{G} = \mathcal{A}$, and the process continues until the finishing criterion is met, i.e., one of the scores $s^* = J(\mathcal{C}, \mathbf{y}^*)$ exceeds the similarity threshold, or the maximum number of iterations is reached.

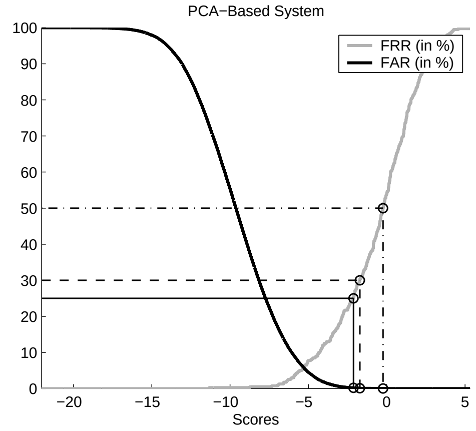

Fig. 1. FAR and FRR curves for the Eigenface-based system.

## III. Face Verification System Attacked

The described Bayesian hill-climbing algorithm is used to attack an Eigenface-based face verification system [9]. This technique uses Principal Component Analysis (PCA) to derive a vector space which represents the face images in a lower dimensional space, this technique was used to present initial face verification results for the recent Face Recognition Grand Challenge [10].

The evaluated system uses cropped face images of size $64 \times 80$ 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 (or eigenvectors). Similarity scores are 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.

## IV. Experimental Protocol

The experiments are carried out on the XM2VTS face database [8], 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 $295 \times 8 = 2,360$ samples of the database. Two evaluation protocols are defined for this database, the Lausanne Protocol 1 and 2 (LP1 and LP2).

### A. Performance Evaluation

The performance of the evaluated system is computed using LP2. The LP2 is chosen as LP1 has a bias to learn the environmental conditions because two independent capture sessions are available whereas for LP2 only one session is available. Also the evaluation data from LP2 is used to compute $\mathcal{G}$, rather than the development data, as there is almost three times the number of impostors available (70 impostors instead of 25); only impostors are used to calculate $\mathcal{G}$ as explained in Sect. IV-B.

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 both the PCA transformation matrix and the enrollment templates), and $ii$) an evaluation set formed by the fourth session images of the previous 200 users (used to compute the genuine scores for the evaluation), and all the 8 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.

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 $\mathcal{C}$. In Fig. 1 we can see the False Acceptance Rate (FAR) and False Rejection Rate (FRR) curves of the Eigenface-based system using the described protocol with $p = 4$ enrollment templates. The system has an Equal Error Rate of 4.74%. 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 [11].

### B. 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 2 sessions of 200 users). The initial $K$-variate distribution $\mathcal{G}$ of the attack algorithm, was estimated using the first session of all the impostors of the test set (70 users) defined in LP2. This way, there is no overlap between the attacked set of users (200 accounts), and the subjects used to initialize the attack 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 $\delta$), divided by the total number of accounts attacked $A_T = 200$. Thus, $SR = A_b/A_T$.

## V. Results

The goal of the experiments is to study the vulnerability of automatic face recognition systems to hill-climbing attacks. At the same time, we will test the viability of applying the Bayesian-based hill-climbing attack algorithm to other biometric traits (it was already successfully applied to attack an on-line signature verification system in [7]).

In the first set of experiments, we study the effect of varying the three parameters of the attack algorithm ($N$, $M$, and $\alpha$) on the success rate (SR) of the attack over the Eigenface-based system working with the standard correlation metric. 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

TABLE I

SUCCESS RATE (IN %) OF THE HILL-CLIMBING ATTACK FOR INCREASING VALUES OF $N$ (NUMBER OF SAMPLED POINTS) AND $M$ (BEST RANKED POINTS). 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 GREY.

| | | $N$ | | | | |
|---|---|---|---|---|---|---|
| | | 10<br>(2500) | 25<br>(1000) | 50<br>(500) | 100<br>(250) | 200<br>(125) |
| $M$ | 3 | 84.5<br>**5,162** | 86.0<br>**4,413** | 86.0<br>**4,669** | 86.0<br>**5,226** | 86.0<br>**6,296** |
| | 5 | 81.5<br>**5,796** | 86.0<br>**4,275** | 86.0<br>**4,512** | 86.0<br>**5,022** | 86.0<br>**5,988** |
| | 10 | | 85.5<br>**4,534** | 86.0<br>**4,540** | 86.0<br>**5,019** | 86.0<br>**5,941** |
| | 25 | | | 86.0<br>**5,213** | 86.0<br>**5,379** | 86.0<br>**6,256** |
| | 50 | | | | 86.0<br>**6,455** | 86.0<br>**6,934** |
| | 100 | | | | | 86.0<br>**8,954** |

the fixed threshold $\delta$. As described in Sect. II, the above mentioned parameters denote: $N$ the number of points sampled from the adapted distribution at a given iteration, $M$ the number of top ranked samples used at each iteration to adapt the global distribution, and $\alpha$ is an adaptation coefficient which varies from $[0 \ldots 1]$. In the last experiment the attack performance is computed for different operating points of the system.

When presenting results the proposed hill-climbing algorithm is compared to a brute-force attack. This provides an overview of the system performance by presenting the number of matches necessary to conduct a successful brute-force attack at the operating point under consideration ($n_{bf} = 1/\text{FAR}$). However, it must be noted that the resource requirements of the brute force attack are much greater because an efficient brute-force attack requires a database of thousands of samples; this is not the case for the hill climbing algorithm.

### *A. Analysis of sampled and retained points*

For the initial evaluation of the algorithm, an operating point of (FRR=50%, FAR=0.01%) was fixed. 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 $\alpha = 0.5$). Results are given in Table I. 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**.

An analysis of the average number of comparisons needed to enter the system given in Table I shows that for $N \gg M$, the points selected to estimate the local distribution are too specific and thus, the attack is slower, needing more comparisons to break the accounts with respect to the best trade-off combination ($N > M$). On the other hand, if $N \simeq M$, the local distribution computed is too general, and again the attack effectiveness in terms of the speed of the attack is reduced. However, this effect cannot be observed in the success rate of the attack, which is 86% for most of the configurations evaluated (172 broken accounts out of a total of 200).

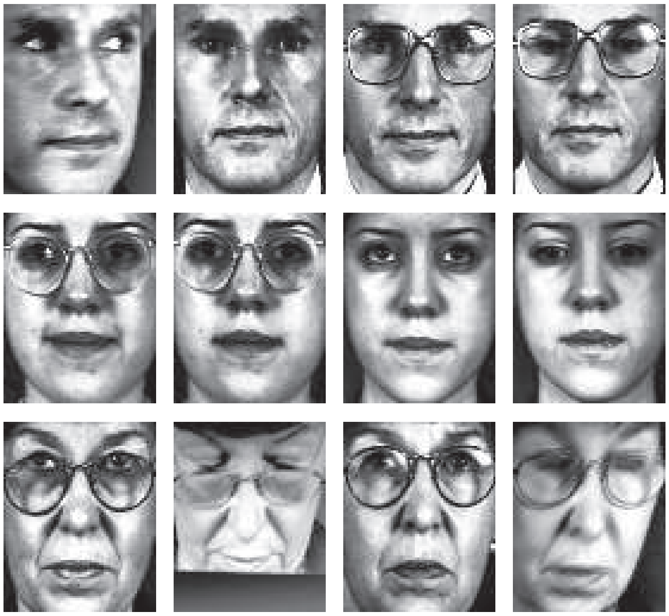

Fig. 2. The four enrollment images (columns) constituting the model of three of the unbroken accounts (rows).

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). None of the client models produced a score high enough to enter the system, which means that these 28 clients would not be suitable for face recognition under the considered system working at the selected operation point. We can then conclude that the attack successfully broke all the models that would be used in a real application. In Fig. 2 the enrollment images which form three of the resistant accounts are showed. In all cases we can observe a great variance among the samples of a given model (glasses/not glasses, different poses, and blurred images).

### *B. Analysis of the adaptation coefficient*

The best configuration of the parameters that can be extracted from Table I corresponds to $[N = 25, M = 5]$. For this point, the number of accounts broken is maximized (86%), and $n_{comp}$ is minimized (4,275). This minimum represents less than half of the expected number of matchings required for a successful brute-force attack ($n_{bf} = 1/\text{FAR} = 10,000$).

For this configuration and for the same operating point (FAR=0.01%), the effect of varying the adaptation coefficient

TABLE II

SUCCESS RATE (IN %) OF THE HILL-CLIMBING ATTACK FOR INCREASING VALUES OF $\alpha$ AND FOR $[N, M] = [25, 5]$. THE SUCCESS RATE (IN %) APPEARS IN PLAIN TEXT, WHILE THE AVERAGE NUMBER OF ITERATIONS NEEDED TO BREAK AN ACCOUNT APPEARS IN **BOLD**.

| $\alpha$ | 0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| SR(%)<br>$n_{comp}$ | 0<br>**25,000** | 84.5<br>**6,468** | 86.0<br>**5,121** | 86.0<br>**4,617** | 86.0<br>**4,381** | 86.0<br>**4,273** | 86.0<br>**4,380** | 81.0<br>**4,990** | 71.5<br>**7,901** | 51<br>**10,404** | 20.0<br>**14,154** |

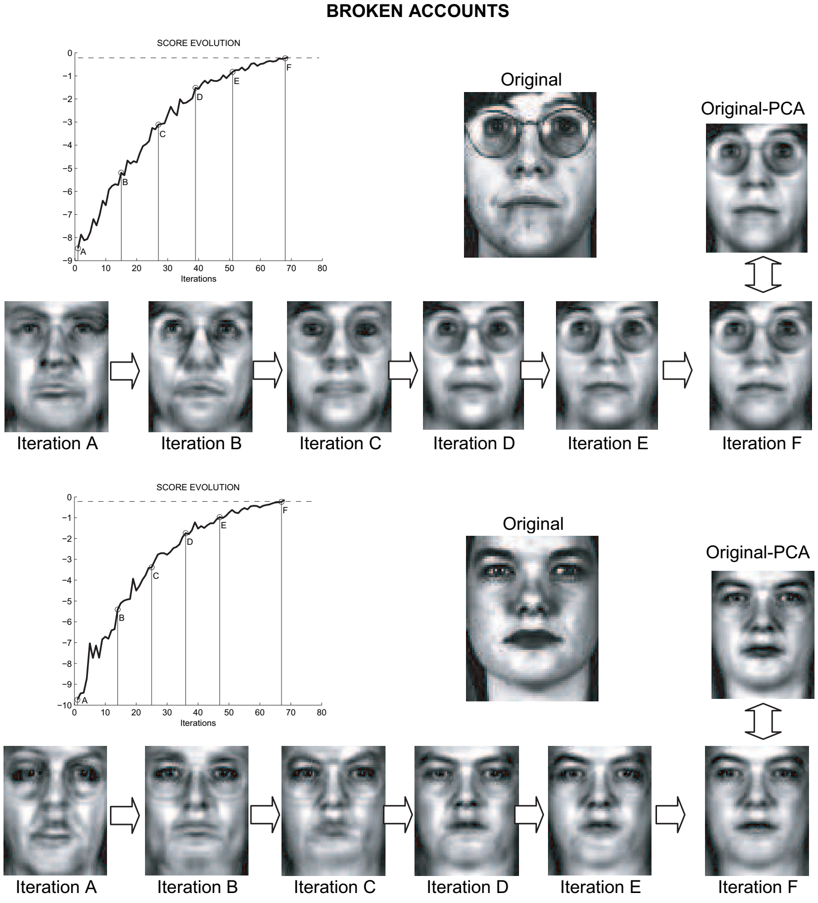

Fig. 3. Examples of the evolution of the score and the synthetic eigenfaces through the iterations of the attack for broken accounts. The dashed line represents the objective threshold.

$\alpha$ on the performance of the attack is studied sweeping its value from 0 (only the global distribution $\mathcal{G}$ is taken into account), to 1 (only the local distribution $\mathcal{L}$ affects the adaptation stage). The results are presented in Table II.

It can be observed that the maximum number of accounts broken, and the minimum number of comparisons needed, is reached for $\alpha = 0.5$, which means that both the global and local distributions should be given approximately the same importance. As in the previous experiment, it can be noticed that 14% percent of the accounts (the same 28 clients as in the previous experiments) is never bypassed as a consequence of the large user intra-variability.

### C. Analysis of different operating points

Using the best configuration $[N, M, \alpha] = [25, 5, 0.5]$ the algorithm was evaluated at two additional operating points of the system (see Fig. 1). The two additional operating points are: $i$) FAR=0.05%, FRR=30% and $ii$) FAR=0.1%, FRR=25%, where a random brute-force attack would need on average $n_{bf} = 2,000$ and $n_{bf} = 1,000$ matches respectively before gaining access to the system. Results are given in Table III.

Smaller values of the FAR rate imply a bigger value of the threshold $\delta$ to be reached by the algorithm, which causes a rise in the average number of iterations required for a successful attack. In any case, the number of comparisons needed to break the system by the hill-climbing algorithm is always lower than that of a brute force attack, with the additional advantage that only a few real face images are needed to initialize the hill-climbing scheme.

In Figs. 3 and 4 two examples of broken and non-broken

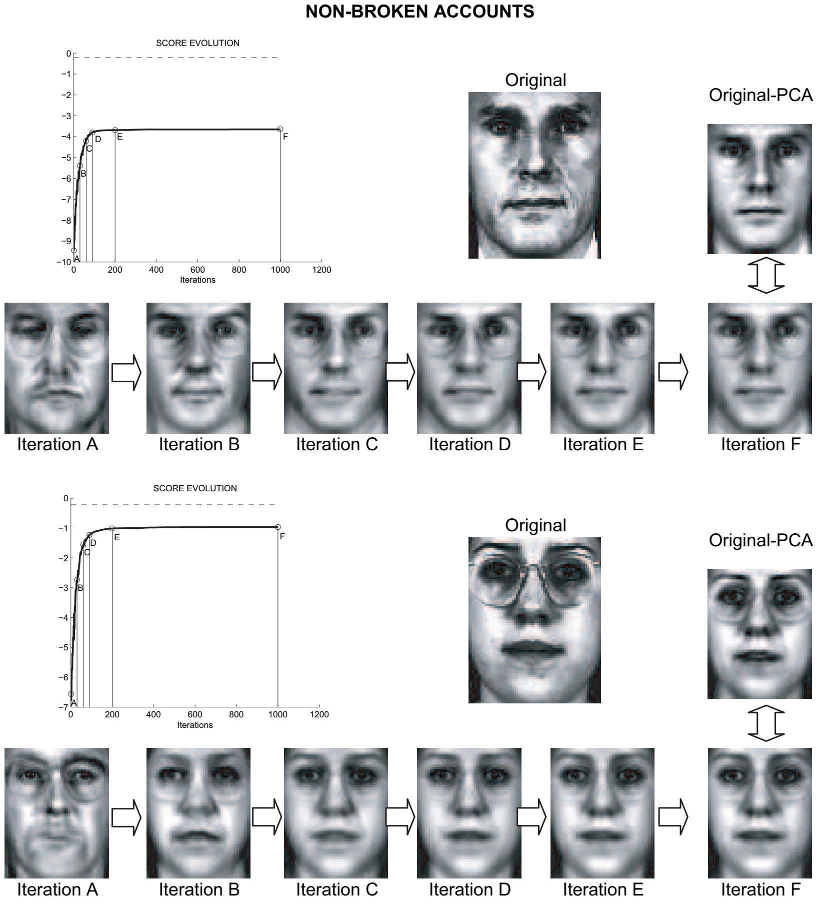

Fig. 4. Examples of the evolution of the score and the synthetic eigenfaces through the iterations of the attack for non-broken accounts. The dashed line represents the objective threshold.

TABLE III
RESULTS OF THE ATTACK FOR DIFFERENT POINTS OF OPERATION AND THE BEST CONFIGURATION FOUND OF THE ATTACKING ALGORITHM ($N = 25$, $M = 5$, $\alpha = 0.5$). 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.

| | Operating points (in %) | | |
|---|---|---|---|
| | FAR=0.1<br>FRR=25 | FAR=0.05<br>FRR=30 | FAR=0.01<br>FRR=50 |
| Success Rate (in %) | 99.0 | 98.5 | 86.0 |
| $n_{comp}$ | **840** | **1,068** | **4,492** |
| $n_{bf}$ | 1,000 | 2,000 | 10,000 |

accounts (corresponding to two of the users presented in Fig. 2) are shown. For each of the examples, the evolution of the score through the iterations of the algorithm is depicted, together with six points (including the first and the last one) of the iterative process (marked with letters A to F). The dashed line represents the objective value to be reached (i.e., the threshold $\delta$). The two upper faces correspond to one of the original images of the attacked user and its representation in the PCA space (where part of the information has been lost). The sequence of the six faces below correspond to the feature vectors that produced each of the six scores marked with A to F, including the first one A which is produced by randomly sampling the estimated general distribution $\mathcal{G}$ and the last one F which is able to break the system.

In Figs. 3 and 4 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"). 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.

## VI. CONCLUSIONS

The robustness of a PCA-based face verification system against a hill-climbing attack based on Bayesian adaptation has been studied. Experimental results show that the system is highly vulnerable to this type of attacking approach which reached a success rate of over 85% in all the experiments executed.

Although the hill-climbing algorithm used in the experiments and a brute force attack are not fully comparable, the performance of both approaches were confronted. The results show that the 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 just needs a small set of samples to initialize the algorithm.

At the same time, it has been proven that the hill-climbing algorithm algorithm can be successfully applied to different biometric traits (in [7] it was tested with very good results on an on-line signature verification system). Thus, this threat should be taken into account when designing any biometric security system working with fixed length feature vectors, and the necessary countermeasures against the attack should be introduced.

## Acknowledgment

J. G. is supported by a FPU Fellowship from the Spanish MEC, and the postdoctoral research of J. F. 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] A. K. Jain, A. Ross, and S. Pankanti, "Biometrics: a tool for information security," *IEEE Trans. on Information Forensics and Security*, vol. 1, no. 2, pp. 125–143, 2006.

[2] P. Mohanty, S. sarkar, and R. Kasturi, "From scores to face templates: a model-based approach," *Pattern Analysis and Machine Intelligence*, vol. 29, pp. 2065–2078, 2007.

[3] C. Soutar, "Biometric system security. http://www.bioscrypt.com/assets/security_soutar.pdf."

[4] A. Adler, "Sample images can be independently restored from face recognition templates," in *Proc. CCECE*, vol. 2, 2003, pp. 1163–1166.

[5] U. Uludag and A. K. Jain, "Attacks on biometric systems: a case study in fingerprints," in *Proc. SPIE-IE*, vol. 5306, no. 4, 2004, pp. 622–633.

[6] 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 *Proc. IEEE ICCST*, 2006, pp. 151–159.

[7] J. Galbally, J. Fierrez, and J. Ortega-Garcia, "Bayesian hill-climbing attack and its application to signature verification," in *Proc. IAPR ICB*. Springer LNCS-4642, 2007, pp. 386–395.

[8] K. Messer, J. Matas *et al.*, "XM2VTSDB: The extended M2VTS database," in *Proc. IAPR AVBPA*, 1999.

[9] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," in *Proc. IEEE CCVPR*, 1991, pp. 586–591.

[10] J. Phillips, P. Flynn *et al.*, "Overview of the face recognition grand challenge," in *Proc. IEEE CCVPR*, 2005, pp. 947–954.

[11] ANSI X9.84-2001, Biometric Information Management and Security.