Robustness of multi-modal biometric verification systems under realistic spoofing attacks

Battista Biggio, Zahid Akthar, Giorgio Fumera, Gian Luca Marcialis, and Fabio Roli
Department of Electrical and Electronic Engineering - University of Cagliari
Piazza d’Armi, 09123 Cagliari, Italy
{battista.biggio, z.momin, fumera, marcialis, roli} diee.unica.it

Abstract

Recent works have shown that multi-modal biometric systems are not robust against spoofing attacks [12, 15, 13]. However, this conclusion has been obtained under the hypothesis of a “worst case” attack, where the attacker is able to replicate perfectly the genuine biometric traits. Aim of this paper is to analyse the robustness of some multi-modal verification systems, combining fingerprint and face biometrics, under realistic spoofing attacks, in order to investigate the validity of the results obtained under the worst-case attack assumption.

1. Introduction

A spoofing attack consists in submitting an artificial replica of a biometric trait to the sensor [1]. For instance, a fake fingerprint or a photo of the targeted client [2, 3, 4]. These replicas are acquired and processed as “live” biometrics. Thus, the verification system may accept them as belonging to a genuine user, if they are very similar to the client’s template.

So far this kind of attacks has been analysed especially for fingerprints [5, 6], but this security issue has been also pointed out for face, iris and other biometrics as well [8, 9, 10]. To address this issue, “liveness” detection methods have been proposed [7]. However, embedding a liveness detection module into a biometric verification system can increase the probability of rejection of a genuine user to an extent that is not acceptable for practical applications.

As an alternative to liveness detection, it has been claimed that multi-modal biometric systems are intrinsically more robust against spoofing attacks [11]. The motivation is that all biometrics should be spoofed to crack a multi-modal system. However, this claim has been recently rejected: it has been shown by experiments that spoofing only one biometric can be sufficient to crack the system, even when more than two biometrics are used [12, 15, 13].

Most of the results reported in [12, 15, 13] refer however to a “worst-case” scenario, obtained by simulating the fake scores under the assumption that their distribution is equal to the genuine users’ one, namely, that the attacker is able to replicate the attacked biometric perfectly. It is thus interesting to further investigate to which extent the above worst-case scenario is realistic, and thus, whether the conclusion drawn from these results (that multi-modal biometrics are not intrinsically more robust against spoofing attacks) holds in realistic scenarios. With regard to this issue, we note that in the works mentioned above one only experiment was carried out in a realistic setting [15], using a small subset of a spoofed fingerprint data set, coming from the LivDet09 Fingerprint Liveness Detection Competition [14]. The corresponding fake score distribution turned out to be significantly different from the genuine one. Quoting from [15], this “indicates that the spoofed fingerprints where not perfect counterfeits in general”. In other words, this result showed that in a realistic case the worst-case assumption may not hold. No experiment was made on realistic spoofed faces instead. In [15] it was argued that the worst-case assumption is reasonable for 2D face systems, as it is easy to show a picture of a genuine user in front of a camera. However, we point out that a scenario in which the attacker can always capture the exact client’s face used as template into the system may be not realistic.

According to the above motivations, it is still necessary to test the robustness of multi-modal biometric systems under various realistic attack scenarios, beside the ones investigated in [12, 13, 15]. Such analysis should allow pointing out to what extent the drop of performance under the worst-case attack scenario is representative of the performance under real spoofing attacks.

To this aim, in this paper we analyse the robustness of multi-modal biometric systems made up of a face and a fingerprint matcher, against spoofing attacks that represent several realistic attack scenarios. To create realistic spoofs, we collected four data sets. Two data sets contain fake fingerprints fabricated with silicon and latex. Two other data
sets contain spoofed faces. One of these two data sets was obtained by showing on a monitor the available testing images of clients, and capturing them with the camera. The other one is obtained by using personal photos voluntarily provided by the clients, to simulate attacks in which photos are collected from the Web, for instance from social networks. These four realistic data sets are combined in order to obtain “chimerical” bi-modal data sets in which either of the biometrics, or both of them, are spoofed. We carried out our experiments using well known score fusion rules: product, LDA-based weighted sum, and Likelihood Ratio (LLR). We also used a modified version of LLR, Extended Likelihood Ratio (ExtLLR), which was proposed in [12] to improve the robustness of LLR against spoofing attacks.

Our experimental results show that the predicted drop of performance under spoofing attacks under the worst case scenario is not always realistic. Accordingly, score fusion rules designed to be robust against spoofing attacks, but depending on the worst-case assumption, as the one proposed in [12], may be not effective in realistic scenarios.

The rest of the paper is organised as follows. Section 2 describes the adopted fusion rules. The data sets are described in Section 3. Experimental results are reported in Section 4. Conclusions are drawn in Section 5.

2. Fusion rules used

The system used in our experiments is shown in Fig. 1. It is a standard bi-modal verification system based on face and fingerprint. When a pair of face and fingerprint images is submitted to the system, they are processed and the related set of features is compared with the one of the system templates related to the claimed identity, by two independent matching algorithms. Result is a pair of matching scores.

The main focus of this paper is on the fusion rule. Several fusion methods may be used: at feature-level, score-level, or decision-level [11]. In this paper we adopted the score-level fusion scheme, as in [12, 13, 15].

In the following, \( s_1 \) and \( s_2 \) denote scores provided respectively by face and fingerprint matchers, and \( s = f(s_1, s_2) \) is the score-level fusion rule. Score-level fusion rules can be subdivided into fixed and trained. The difference is that the former include a set of parameters to be estimated from training data. Finally, an acceptance threshold \( s^* \) must be set such that, if \( s \geq s^* \), the claimed identity is accepted and the person is classified as a genuine user, otherwise the it is classified as an impostor.

In this work, we used one fixed and three trained rules.

**Simple product.** It is a simple fixed rule where:

\[
s = s_1 \cdot s_2. \tag{1}
\]

**Weighted Sum by Linear Discriminant Analysis (LDA).** This is a trained rule where individual matching scores are linearly combined:

\[
s = w_0 + w_1 s_1 + w_2 s_2. \tag{2}
\]

Weights \( w_0, w_1 \) and \( w_2 \) are set according to the maximisation of the Fisher distance \( FD \) between the score distributions of genuine and impostor users. In the case of two matchers, \( FD \) is defined as follows:

\[
FD = \frac{(\mu_1 + \mu_C)^2}{\sigma_1^2 + \sigma_C^2}, \tag{3}
\]

where \( \mu_1 \) and \( \mu_C \) are the means respectively of the impostor and genuine score distributions, while \( \sigma_1^2 \) and \( \sigma_C^2 \) are their variances.

**Likelihood Ratio (LLR).** This is the so-called Neyman-Pearson test:

\[
s = \frac{p(s_1, s_2|G)}{p(s_1|G) \cdot p(s_2|G)} = \frac{p(s_1|G) \cdot p(s_2|G)}{p(s_1|I) \cdot p(s_2|I)}, \tag{4}
\]

where a conditional independence between \( s_1 \) and \( s_2 \), given that they come either from an impostor or a genuine user, is often assumed. While in this case \( s \) is not a matching score, since it ranges in \([0, \infty)\) (instead of \([0, 1]\) as usual), the final classification scheme is the same as above.

**Extended LLR (ExtLLR).** This is a variant of the LLR, and was proposed in [12] to make LLR robust against spoofing attacks. The basic idea was to explicitly take into account the probability distribution of spoof attacks when modelling the probability distribution of the impostor class. In this rule, the following parameters were considered besides likelihood of genuine users and impostors: \( \alpha \), which is the probability of a spoofing attempt, and a variable \( c \), for each of the biometrics, representing the probability that, given that a spoofing attempt is made against the corresponding biometric, it does not succeed. It was further assumed that the fake score distribution of successful spoof attempts against any of the biometrics is identical to the corresponding genuine score distribution (the so-called worst-case assumption), and that the one of unsuccessful attempts equals the impostor distribution. The resulting expression of the joint likelihood \( p(s_1, s_2|I) \) for a bi-modal system is:

\[
p(s_1, s_2|I) = \frac{\alpha}{\alpha + (1 - \alpha)(1 + c_1)(1 + c_2)p(s_1|G)p(s_2|G)} + \frac{\alpha}{\alpha + (1 + c_1)(1 - c_2)p(s_1|I)p(s_2|G)} + \frac{\alpha}{\alpha + (1 - c_1)(1 - c_2)p(s_1|G)p(s_2|G)} + \frac{\alpha}{\alpha + (1 - \alpha) + \frac{\alpha}{\alpha + (c_1 + c_2 + c_1c_2)}p(s_1|I)p(s_2|I)},
\]
where the first two summands are related to a successful spoofing attempt against either of the biometrics, the third one corresponds to a successful spoofing attempt against both biometrics, and the fourth one accounts for all unsuccessful spoof attempts, and to standard impostor attacks without spoofing attacks (see [12] for further details).

LLR and ExtLLR require the estimation of individual likelihood. To this aim, we fit the available genuine and impostor matching scores with a parametric distribution [18, 19, 12]. A Gamma distribution was used, as done in [12], because it turned out to provide a good approximation of our data. Moreover, the parameters \( \alpha, c_1 \) and \( c_2 \) were set to the same values used in [12], respectively 0.01, 0.3, and 0.7. Note that in a real application the values of these parameters, namely the probability of a spoofing attempt, and the probability that a spoofing attempt against each of the considered biometrics does not succeed, can not be estimated from training data, and can only be hypothesised.

### 3. Spoof data sets

The size and the characteristics of the data sets described in the following sections are reported in Table 1.

#### 3.1. Face spoofs

Face biometric spoof attack, also known as “copy attack”, is a serious security issue for face recognition systems. In general, face recognition systems could be spoofed by (i) a photo of a genuine user; (ii) a video of a genuine user; (iii) a 3D face model of a genuine user.

The most common, cheapest and easiest face spoofing attack is to submit a photo of a genuine user to the face recognition systems. Since face is not secret, like other biometric traits, it can be captured easily by a camera without the knowledge and consent of the genuine user. Also, due to intertwined network society culture, personal facial photographs are usually accessible to the public. For instance, an impostor can obtain the photographs of genuine users from a social network, and submit them to a biometric authentication system to fool it. Face video of a genuine user can also be easily captured by tiny cameras for spoofing, which later can be presented to a system using a portable device. This type of face spoofing has physiological clues such as facial expression, head movement, and blinking, that may also defeat liveness detection methods based on these clues. 3D face model spoofing is not easy and requires more efforts to imitate the genuine users. Hence, photo and video are the most common faking methods that can be used to spoof face recognition systems.

Since no face biometric data sets including spoofed samples are publicly available, we collected and built a face data set including two kinds of spoofed faces: the Photo Attack and the Personal Photo Attack data sets. To this aim, we collected the live face data set in two sessions, with a time interval of about two weeks between them, under different lighting conditions and facial expressions.

We created the Photo Attack data set using the “photo attack” method described in [4, 8]. It consists in displaying a photo of the targeted client on a laptop screen, which is then put in front of the camera. To this aim, the testing images of clients of the live face data set were used. This simulates a scenario in which the attacker can obtain photos of the targeted client under a setting similar to the one of the verification phase.

To build the Personal Photo Attack data set of spoofed faces, we used personal photos voluntarily provided by 25 clients of the live face data set. On average, we collected 5 photos per client. These photos were taken in different times and under different environmental conditions than those of the live templates. This simulates a scenario where the attacker uses a photo of the targeted client taken from the Web, for instance from social networks.

Fig. 2 shows an example of the original template image of one of the clients, a spoof obtained by the photo attack, and a spoof obtained from an image voluntarily provided by the same client. These two spoofs reflect two different degrees of expected effectiveness, but also of realism. In fact, a photo attack based on one of the images in the data set appears to have, by visual inspection, more chances to be successful than a spoof obtained by personal photos, as the latter may be images significantly different from the ones used as templates by a biometric system. On the other hand, it is more likely that an attacker can obtain a photo of the targeted client from the Web. Consequently, we may expect that the fake score distribution of our Photo Attack data set is close to the one of genuine users, whilst the effectiveness of a spoof attack based on personal photos strongly depends on the ability of the attacker to obtain images similar to the templates used by the system.

#### 3.2. Fingerprint spoofs

Several methods to spoof fingerprints have been proposed, since the early papers of [2, 3]. The most adopted approach is the so-called “consensual” method, which is

<table>
<thead>
<tr>
<th>Data set</th>
<th>Number of clients</th>
<th>Number of spoofs per client</th>
<th>Number of live per client</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latex</td>
<td>80</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>Silicon</td>
<td>142</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>Photo Attack</td>
<td>40</td>
<td>60</td>
<td>60</td>
</tr>
<tr>
<td>Personal Photo Attack</td>
<td>25</td>
<td>3 (avg.)</td>
<td>60</td>
</tr>
</tbody>
</table>

Table 1. Characteristics of the fake fingerprint and fake face data sets used in the experiments.
made up of the following steps:

1. The user puts his finger on a soft material (Play Doh, dental impression material, plaster, etc.).
2. The negative impression of the fingerprint is fixed on the surface.
3. A mould is formed. Silicone liquid or another similar material (wax, gelatine, etc.) is poured in the mould.
4. When the liquid is hardened, the stamp is formed.

We followed the same procedure above to build our spoof data sets. As the mold was made up of plasticine-like material, spoofs were made up of three basic materials: gelatine, silicon, and latex. These three materials are commonly adopted for replicating fingerprints, and have been used for assessing the performance of fingerprint liveness detection systems at First International Competition on Fingerprint Liveness Detection (LivDet09) [14].

Fingerprint sensor adopted is the well-known Biometrika FX2000. For sake of space, in this paper we investigate the robustness of multi-modal systems using only two materials for fabricating spoofs, namely, silicon and latex. Some images, showing the average quality of provided spoofs, are shown in Fig. 3. This figure shows the original, “live” client image, beside a replica made up of latex, and a replica made up of silicon. As it can be seen, the latex image is very similar to the original one, whilst the second one is also characterised by some artefacts. These two replicas, in our opinion, show a reasonable set of realistic scenarios.

4. Experimental Results

We used a similar experimental protocol as in [12, 13]:

- Due to the absence of multi-modal data sets including spoofing attacks, we built four chimerical data sets, by randomly associating face and fingerprint images of pairs of clients of the available face (Photo Attack and Personal Photo Attack) and fingerprint (latex and silicon spoofs) data sets. Note that building chimerical data sets is a widely used approach in experimental investigations on multi-modal biometrics (see [11]).
  - To carry out more runs of the experiments, each of the four data sets was randomly subdivided into five pairs of training and testing sets. Furthermore, all the above procedure was repeated five times, for different random associations of face and fingerprint images of pairs of clients. In each run, the parameters of the trained fusion rules have been estimated on the training set. The results reported below refer to the average test set performance, over the resulting twenty-five runs.
  - The fake match scores were computed by comparing each fake image of a given client with the corresponding template image.
  - The performance was assessed by computing DET curves (FAR vs. FRR). Note that, in the evaluation of spoofing attacks, the FAR corresponds to the percentage of spoofing attempts that were accepted as genuine (the term SFAR was used to this aim in [13]).

The NIST Bozorth3 matching algorithm was used for fingerprint verification [16]. It is based on matching the fingerprint minute details, called “minutiae”. The Elastic Bunch Graph Matching (EBGM) algorithm was used for face verification [17]. It is based on representing a face with a graph whose nodes are the so-called face “landmarks” (centred on the nose, eyes, and other points detected on the face), are labelled by a feature vector, and are connected by edges representing geometrical relationships among them.

We investigated three attack scenarios: (a) only fingerprints are spoofed; (b) only faces are spoofed; (c) both fingerprints and faces are spoofed (bi-modal or double spoofing). For the scenarios (a) and (b) we also evaluated a worst-case attack as defined in [12, 15, 13]: we generated fictitious fake scores by randomly drawing scores of genuine users.
The results are reported in Fig. 4. For the sake of space, we report only the results of two out of the four chimerical data sets used in the experiments: latex spoofed fingerprints and photo attack spoofed faces (Fig. 4, top row); silicon spoofed fingerprints and personal photo attack spoofed faces (Fig. 4, bottom row). Similar results were obtained in the other two data sets. Each column of Fig. 4 refers to a different score fusion rule.

From the top row of Fig. 4, it is easy to see that the worst-case assumption is realistic when faces are spoofed by a photo attack, using an image similar to the template: the fake score distributions are very close to the ones of genuine users. Thus, modelling fake score distributions as genuine ones, as proposed in [15], seems acceptable in this scenario. The same does not hold however for latent-based fake fingerprints: as can be seen by the first three plots in the top row, the corresponding FAR was clearly overestimated by the worst-case assumption (being equal the FRR), with the only exception of the Extended LLR rule.

In the bottom row of Fig. 4, the differences between the FAR attained under the worst-case assumption, and the one observed on our data sets, is even higher, both for spoofed faces and for spoofed fingerprints. In the case of face spoofing, the performance is very close to the one attained without a spoof attack. In the case of fingerprint spoofing, the performance is considerably far both from the one attained in the worst-case scenario, and the one attained without spoofing attacks.

Let us now compare the different fusion rules used in these experiments. Both in the case of spoofing attacks against only one biometric (either face or fingerprint), and of double spoofing, standard rules (Product, LDA and LLR) did not exhibit appreciable performance differences. In other words, they exhibited a similar robustness. Notably, the Extended LLR exhibited a significantly worse performance than standard rules, despite it was specifically designed to be robust under spoofing attacks. This behaviour seems due to the fact that the worst-case assumption behind this rule turned out to be too pessimistic. Note also that, as pointed out in Sect. 2, another problem of Extended LLR is that setting its parameters ($c_1$, $c_2$, and $\alpha$ for a bi-modal system) is not easy, and their values can only be hypothesised.

The above results clearly point out that the worst-case assumption is not a suitable one to assess the robustness of a multi-modal system against spoof attacks, as well as to design fusion rules robust against spoofing attacks. They also suggest that a more realistic modelling of the fake score distribution against spoof attacks is needed to this aim.

5. Conclusions

To assess the robustness of a multi-modal biometric system under spoofing attacks, previous works assumed a worst-case scenario in which the attacker is able to replicate perfectly the genuine biometric traits, and thus the fake score distribution is identical to the one of genuine users [12, 13]. This lead to the conclusion that multi-modal biometric systems are not intrinsically robust against spoofing attacks, as they can be cracked by spoofing only one biometric. A score fusion rule specifically designed to be robust against spoofing attacks was also proposed in [12], based on the above worst-case assumption.

However, through experiments carried out on several realistic spoofing attacks against face and fingerprints, and using four different score fusion rules (including the one of [12]), we provided evidence that a worst-case scenario can be not representative of realistic spoofing attacks. In particular, we found that the above worst-case assumption can be too pessimistic, resulting in a significant overestimation of the FAR that a multi-modal system may incur under a realistic spoofing attack. This can also undermine the effectiveness of score fusion rules based on such assumption, like the one of [12], that turned out to be less effective than standard rules like Product, LDA and LLR.

Our results suggest that an interesting open issue to investigate is the development of more realistic models of the fake score distribution under spoofing attacks, aimed both at assessing the robustness of multi-modal systems, and to design novel, robust score fusion rules.

Acknowledgment

This work was partly supported by the TABULA RASA project, 7th Framework Research Programme of the European Union (EU), grant agreement number: 257289; by the PRIN 2008 project “Biometric Guards - Electronic guards for protection and security of biometric systems” funded by the Italian Ministry of University and Scientific Research (MIUR); and by a grant awarded to B. Biggio by Regione Autonoma della Sardegna, PO Sardegna FSE 2007-2013, L.R. 7/2007 “Promotion of the scientific research and technological innovation in Sardinia”.

References


Figure 4. Average DET curves attained on the test set using latex spoofed fingerprints and photo attack spoofed faces (top), and using silicon spoofed fingerprints and personal photo attack spoofed faces (bottom). Each column refers to a different score fusion rule, indicated in the title of each plot. Each plot contains the DET curves attained with no spoofing attack (black), under realistic spoofing attacks (solid curves), and under simulated worst-case spoofing attacks (dashed curves). Red: fingerprint spoofing only. Blue: face spoofing only. Green: both face and fingerprint spoofing.

<table>
<thead>
<tr>
<th>Product</th>
<th>LDA</th>
<th>LLR</th>
<th>Extended LLR</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>FAR (%)</th>
<th>FRR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10^-1</td>
<td>10^0</td>
</tr>
<tr>
<td>10^-2</td>
<td>10^1</td>
</tr>
<tr>
<td>10^-3</td>
<td>10^2</td>
</tr>
<tr>
<td>10^-4</td>
<td>10^3</td>
</tr>
</tbody>
</table>


[14] http://prag.diee.unica.it/LivDet09 1, 4


