Security Evaluation of Biometric Authentication Systems Under Real Spoofing Attacks

Battista Biggio, Zahid Akhtar, 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

Multimodal biometric systems are commonly believed to be more robust to spoofing attacks than unimodal systems, as they combine information coming from different biometric traits. Recent work has shown that multimodal systems can be misled by an impostor even by spoofing only one biometric trait. This result was obtained under a "worst-case" scenario, by assuming that the distribution of fake scores is identical to that of genuine scores (i.e., the attacker is assumed to be able to perfectly replicate a genuine biometric trait). This assumption also allows one to evaluate the robustness of score fusion rules against spoofing attacks, and to design robust fusion rules, without the need of actually fabricating spoofing attacks. However, whether and to what extent the "worst-case" scenario is representative of real spoofing attacks is still an open issue. In this paper, we address this issue by an experimental investigation carried out on several data sets including real spoofing attacks, related to a multimodal verification system based on face and fingerprint biometrics. On the one hand, our results confirms that multimodal systems are vulnerable to attacks against a single biometric trait. On the other hand, they show that the "worst-case" scenario can be too pessimistic. This can lead to too conservative choices, if the "worst-case" assumption is used for designing a robust multimodal system. Therefore, developing methods for evaluating the robustness of multimodal systems against spoofing attacks, and for designing robust ones, remains a very relevant open issue.

1 Introduction

In the past few years, potential vulnerabilities of biometric systems and related attacks have been detected. Some works have revealed that not only individual modules of a biometric system can be attacked, but also the channel connecting them [27, 7, 32, 17]. Besides the so-called "indirect attacks", which require some knowledge of the system or access to its internal parameters [32, 1, 13, 24], much more attention has been devoted to the so-called "direct attacks", namely, attacks which consist of submitting a counterfeited biometric (i.e., a replica of the client's biometric)

to the sensor. They are mostly referred to as *spoofing attacks* [27, 32, 16, 14, 6]. The effectiveness of these attacks relies upon the fact that the biometric sensor is unable to detect whether the submitted biometric trait is "fake" (i.e., "spoofed") or "live", and, thus, it generates an image which is then processed and potentially recognized as a "live", genuine biometric trait.[1] Therefore, carefully counterfeited biometrics represent a serious danger both for biometric identification and verification systems.

In this paper, we focus on biometric identity verification. In this kind of task, the user claims the identity of an enrolled client, and provides his biometric traits to the system. The claimed identity is then verified by comparing the submitted biometric traits with those stored into the system database corresponding to the claimed identity (which are usually referred to as "templates"). The output of this step is a matching score which is then compared with a decision threshold: if the matching score is above the threshold, the user is accepted as *genuine*; otherwise, he is rejected as *impostor*.

Although there are threshold values that, theoretically, should allow the system to avoid impostor acceptance (e.g., the so-called zeroFAR operating point), it is also true that the protocols used to set these operating points rely on the hypothesis that the attacker's biometric is intrinsically different from that of the targeted clients, and that such hypothesis is likely to be violated in spoofing attacks, as the attacker tries to impersonate a specific client through the replication of his biometric traits [19].

The issue of spoofing attacks has been especially studied in the case of fingerprints and faces, and, rarely, on other biometrics, in the novel research field of *biometric liveness detection* [11, 20, 9, 10, 23, 33, 22, 21, 34, 4, 6]. Several hardware- and software-based liveness detection systems have been proposed (see, e.g., [9]), but neither the first ones, nor the second ones, have shown acceptable performances and costs against direct attacks.

Until a few years ago, it was commonly believed that multimodal biometric systems were not only able to outperform unimodal systems in terms of classification performance, but also that they were intrinsically robust to direct attacks. This claim was based on the intuition that more than one biometric trait should be spoofed to mislead these systems, although it was not supported by experimental evidences or theoretical findings. In fact, conversely to the above claim, the first experimental evidences showed that multimodal systems can be deceived with high probability by attacking only one of the available biometric modalities [29, 19]. However, this result was derived without fabricating any *real* fake trait, and only simulating the corresponding matching scores, under the hypothesis of "perfect" duplication of the targeted biometric. In particular, it was assumed that the distribution of the matching scores of the fake traits was identical to that of the "live" traits of genuine users. Since this amounts to assuming that the matcher under attack outputs the same scores for genuine users and impostors attempting a spoofing attack (namely, it would accept both of them with the same probability), we will refer to this hypothesis as "worst-case".

As this assumption allows one to simulate spoofing attacks without the need of fabricating any fake biometric trait, a system designer may easily evaluate the performance drop of different fusion rules under direct attacks, and detect

[1] In this paper, we will use both "fake" and "spoofed" terms interchangeably, to indicate an artificial replica of the client's biometric.

the most robust fusion rule under the considered attack scenario. This would eventually lead to better design choices, as both classification performance and robustness to spoofing attacks can be exploited as decision criteria. Moreover, the "worst-case" assumption can also be exploited to improve robustness of fusion rules to spoofing attacks, or to design novel and more robust fusion rules. For instance, to this end, a modification of the well-known likelihood ratio score fusion rule was proposed in [29].

Besides this, whether and to what extent the above "worst-case" scenario is representative of real spoofing attacks is still an open issue. In fact, this hypothesis is only supported by a very limited experimental evidence reported in [28], where the distributions of some real spoofing attacks were shown to be similar to those of the "live" traits of genuine users. A more systematic and wider experimental analysis is however still required.

To this end, in this paper, we extended our preliminary analysis in [5] by reporting a larger set of experiments on real data sets of spoofing attacks, against a bimodal system made up of fingerprint and face biometrics. We considered several data sets consisting of real spoofs, for both fingerprints and faces. Fake fingerprints were fabricated using several materials, leading to a very large set of spoofs, which can be retained representative of the state-of-the-art. This is witnessed by the fact that these spoofs were also used in the First and Second Fingerprint Liveness Detection Competition [23, 33]. With regard to fake faces, we considered different attack scenarios depending on how the attacker fabricates his fakes, and built the corresponding data sets. We also considered the public data set recently used in the Competition on Countermeasures to 2D Facial Spoofing Attacks [4, 6]. Our experiments were carried out on a large number of fusion rules, including the one proposed in [29], which is explicitly designed to be robust to spoofing attacks. We also simulated "worst-case" spoofing attacks targeting each biometric trait separately, based on the same hypothesis in [29, 19, 28], in order to compare the corresponding results with those attained by real attacks targeting the same trait, and also with real attacks targeting both biometrics.

Our results show that, while the "worst-case" scenario is often representative of spoofing attacks against face matchers, it turns out to be too pessimistic in the case of fingerprints, even when high quality fakes are used. As a consequence, using the "worst-case" assumption can be unsuitable to compare the robustness of different score fusion rules against spofing attacks. Moreover, score fusion rules specifically designed to take into account spoofing attacks and based on the "worst-case" scenario, like the one proposed in [29], can be ineffective or even harmful, when such scenario is not representative of the actual fake score distribution. These remain therefore very relevant research issues to be addressed in future work.

The paper is organized as follows: Sect. 2 gives an overview of spoofing attacks, for both fingerprint and face biometrics; Sect. 3 summarizes multimodal biometric systems, and some well-known and widely used fusion rules; Sect. 4 describes our experiments, and discusses the achieved results; and Sect. 5 draws the conclusions and sketches future work.

2 Spoofing attacks

As mentioned in Sect. 1, a spoofing attack consists of stealing, copying and replicating synthetically a biometric trait, to gain unauthorized access, defeating the biometric system security [16, 14, 6]. The feasibility of a spoofing attack is much higher than other types of attacks against biometric systems, as it does not require any knowledge on the system, such as the feature extraction or matching algorithm used. Digital protection techniques like hashing, encryption, and digital signature, are not useful due to the nature of spoofing attacks, which are done in the analogical domain, outside the digital limits of the system. Although liveness detection can be exploited to counteract spoofing attacks, as mentioned in the previous section, state-of-the-art techniques may increase the percentage of genuine users rejected by the system (i.e., the false rejection rate, FRR) [9, 22, 23, 33, 4, 6].

2.1 Fingerprint spoofing

Fingerprint spoofing is quite an old exercise. Alert Wehde carried out the very first endeavor to spoof fingerprints in the 1920s [8]. He, then an inmate at a Kansas penitentiary, used his expertise in photography and engraving to produce a gummy fingerprint from a latent fingerprint. The latent fingerprint was highlighted using forensic methods, and a photograph was taken. The photograph was then used to etch the print onto a copper plate, which was later used to generate fake latent fingerprints on surfaces.

In recent years, several research studies have been conducted to investigate how spoofed fingerprints can circumvent state-of-the-art fingerprint recognition systems. Authors in [26] studied the susceptibility of different biometric fingerprint sensors to fake fingerprints synthesized with silicone and plasticine. Results on six optical and solid-state commercial sensors were reported. Five sensors permitted the unauthorized access into the system on the first attempt, while the remaining one was spoofed on the second attempt.

Matsumoto et al. [25] reported, by conducting similar experiments as in [26], that fake fingerprints fabricated with gelatin are more effective. The authors tested eleven commercial fingerprint sensors, with a success rate higher than 60%, even also when the fake fingerprints were replicated from the latent fingerprint. A similar robustness evaluation of different sensors under fake fingerprints fabricated with several spoofing techniques can be found in [31, 20]. In [20], the authors extended the experiments reported in [25] to test new sensors embedded with fake detection measures. The authors concluded that such fake detection measures were able to reject spoofed fingerprints replicated using non-conductive materials such as silicone, while were not able to detect fake fingerprints fabricated using conductive materials like gelatin.

In [12], the possibility of fabricating fake fingerprints from standard minutiae templates was studied. A two-stage process was carried out to create the spoofed fingerprints. In the first stage, fingerprint images were reconstructed from the genuine user's minutiae template. This stage was termed as "from the template to the image". In the second stage, called "from the image to the gummy fingerprint", the reconstructed images were utilized to produce fake fingerprints.

In spite of some errors that were accumulated during the reconstruction process, more than 70% of the fake fingerprints were accepted by the system.

The existing literature, as described above, suggests that fingerprint spoofing methods can be classified into two broad categories: "consensual/cooperative/direct casts" and "non-consensual/non-cooperative/indirect casts". In the consensual method, the fake fingerprints are created with the consent and collaboration of the fingerprint owner. In the non-consensual method, the latent finger-marks, that the user has unnoticeably left on some surface, are used to fabricate the spoofed fingerprint using a very similar procedure to that mentioned in [8]; hence, the cooperation of the user is not required. It is worth noting that most of the research studies have been conducted using spoofed fingerprints fabricated with the consensual method, that consists of the following steps:

1. the user presses his finger on a soft material such as wax, play doh, dental impression material, or plaster, to create the negative impression of the fingerprint as a mold;
2. a casting material such as liquid silicon, wax, gelatin, moldable plastic, plaster or clay, is poured in the mould;
3. when the liquid is hardened, the fake fingerprint is formed.

In this paper, we followed the consensual method to create high quality fake fingerprints, as done in the First and Second Editions of the International Fingerprint Liveness Detection Competition [23, 33].

2.2 Face spoofing

In spite of the fair amount of advancement in biometric face recognition systems, face spoofing (also known as "copy attack") still poses a serious threat to the system security. Face spoofing methods may vary according to the targeted face recognition system. Face recognition systems can be broadly classified into two groups: 2D (two-dimensional) and 3D (three-dimensional) systems. The former process two-dimensional face images, while the latter extract some features from the 3D shape of faces using ad hoc methods like paraxial viewing, or pattern illumination light [15]. Conventionally, face recognition systems can be spoofed by presenting (i) a photograph, (ii) a video, or (iii) a 3D face model or mask of a genuine user.

Face spoofing through photograph or video is the most common, cheapest and easiest method to circumvent face recognition systems [6, 34]. Spoofing attacks through photograph, known as "photo-attacks", consist of submitting a photograph of a legitimate user to the face recognition system, displayed in hard copy or on the screen of a portable computer or smart phone [6, 34]. Since face is not concealed like other biometrics, it may be easier to spoof; for instance, distant cameras can easily photograph a user's face without the knowledge or prior consent of the user. Furthermore, due to social image sharing websites and social networking websites, many users' photographs can be easily downloaded by an attacker to fool a face recognition system. It is worth noting that the "photo-attack" method

was used to assess the performance of face liveness detection systems at the competition on counter measures to 2D facial spoofing attacks, held in conjunction with the 2012 International Joint Conference on Biometrics [6].

The advent of public video sharing websites and reduction in the cost of high quality video cameras and portable devices have also made it easy to obtain or capture genuine user's facial video samples without subject's consent and awareness, which can be later presented to the system using a portable device for spoofing purpose [34, 6, 21]. The likelihood of success of a video attack becomes higher due to the liveness appearance in the displayed fake faces. For example, a fake face image could be characterized by a facial expression, and head movement (artificially provided by the attacker during the biometric submission), which could thwart liveness detection methods based on these features. It is clearly more difficult to obtain a fake eye blinking with a fake face image (although this may happen by properly tilting the image), whereas all the mentioned features can be easily cheated through a video attack.

Due to the wide adoption of 2D systems across the globe, "photo attacks" and "video attacks" are still the most common techniques to spoof faces. 3D face recognition systems can be spoofed by 3D face models or face masks counterfeited by silicon gel, rubber, or plastic [34]. Alternatively, a 2D photo of the targeted client can be attached to a shirt, and appropriately worn by the attacker on his own face, in order to create a sort of "3D effect". Although it may be easier to attempt this kind of attack than producing a 3D mask, its probability of succeeding might be lower, as the attacker is required to mimic the 3D shapes of the spoofed face by properly adjusting the shirt over his face. In principle, 3D spoofing may also be exploited to defeat 2D face recognition systems, as it might be able to withstand certain 2D anti-spoofing measures. However, no experimental evidence for these cases has been reported thus far.

3 Multimodal biometric systems against spoofing attacks

In this section we first summarize background concepts about multimodal biometric systems, and then focus on previous works that addressed the issue of their robustness to spoofing attacks.

3.1 Background concepts

Multimodal biometric systems have been originally proposed to improve the personal identity recognition performance, through the combination of information coming from different biometric traits, which can overcome the limits and the weaknesses inherent in every individual trait.

A multimodal biometric system is made up of two or more sensors, each one based on a different biometric trait. The information coming from the sensors can be integrated at different levels: sensor, feature, matching score, and decision [30]. In this work, we only one consider fusion at the matching score level, since it is the most commonly adopted approach, and is the one considered in [29, 19, 28]. Without loss of generality, in the following we focus on a system made up of a fingerprint and a face matcher, since it is the one used in the experiments of Sect. 4.

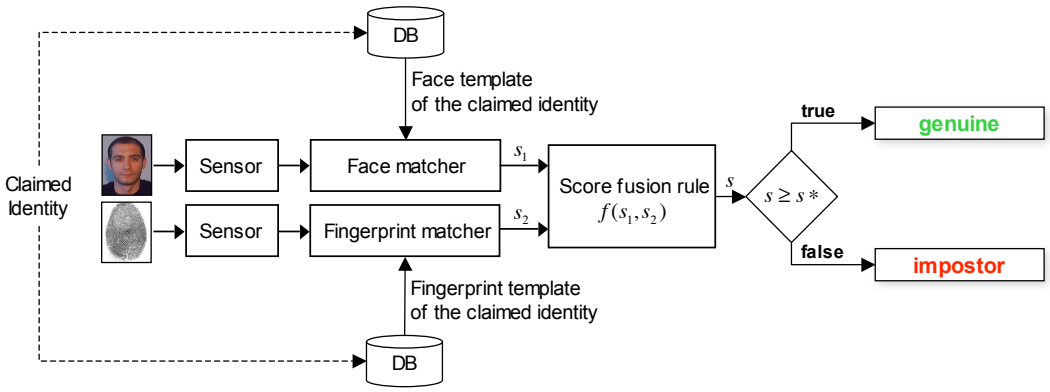

Figure 1: A bimodal biometric system made up of a fingerprint and a face sensor, whose matching scores are combined through a fusion rule.

Such a multimodal systems operates as follows (see also Fig. 1). At the design phase, authorized users (clients) are enrolled: their biometric traits are stored in a database, together with the corresponding identities. During operation, each user submits the requested biometric traits to the sensors, and claims the identity of a client. Then, each matcher compares the submitted trait with the corresponding template of the claimed identity, and provides a real-valued matching score (denoted here as s_1 and s_2, respectively for the fingerprint and the face matcher): the higher the score, the higher the similarity. Finally, the matching scores are combined through a fusion rule which outputs a new real-valued score $f(s_1, s_2)$: the claimed identity is accepted and the person is classified as a genuine user, if $f(s_1, s_2) \geq s^*$; otherwise, it is classified as an impostor. The term s^* is an acceptance threshold that must be set during design according to application requirements in terms of false acceptance (FAR) and false rejection (FRR) rates.

Score-level fusion rules can be subdivided into fixed and trained. The difference between them is that the latter include a set of parameters to be estimated from training data. We describe here the most widely used rules, which will be used in the experiments of Sect. 4.

Sum. It is a simple fixed rule defined as:

$$f(s_1, s_2) = s_1 + s_2 \ . \tag{1}$$

Product. It is another fixed rule defined as:

$$f(s_1, s_2) = s_1 \times s_2 \ . \tag{2}$$

Weighted sum by Linear Discriminant Analysis (LDA). This is a trained rule which consists of linearly com-

bining the individual matching scores:

$$f(s_1, s_2) = w_0 + w_1 s_1 + w_2 s_2 \ . \tag{3}$$

The weights w_0, w_1 and w_2 are set as the ones that maximize the Fisher distance (FD) between the score distributions of genuine and impostor users. In the case of two matchers, FD is defined as follows:

$$FD = \frac{(\mu_{\mathrm{G}} - \mu_{\mathrm{I}})^2}{\sigma_{\mathrm{G}}^2 + \sigma_{\mathrm{I}}^2} \ , \tag{4}$$

where μ_{G} and μ_{I} are the means of the genuine and impostor score distributions, respectively, and σ_{G}^2 and σ_{I}^2 are their variances.

Likelihood ratio (LLR). This is a trained rule which corresponds to the so-called Neyman-Pearson test:

$$f(s_1, s_2) = \frac{p(s_1, s_2|\mathrm{G})}{p(s_1, s_2|\mathrm{I})} \ . \tag{5}$$

Conditional independence between s_1 and s_2, given that they come either from an impostor or a genuine user, is often assumed, so that $p(s_1, s_2|\cdot) = p(s_1|\cdot)p(s_2|\cdot)$. Note that, in this case, $f(s_1, s_2)$ is not obtained as a matching score between two biometric traits, but as a ratio between likelihoods. Nevertheless, the decision rule is the same as above.

3.2 Robustness against spoofing attacks

Besides being more accurate than biometric systems based on a single modality, multimodal systems are commonly believed to be more robust to spoofing attacks. Such belief is based on the intuitive argument that, to evade a multimodal system, the attacker needs to spoof *all* the corresponding biometric traits simultaneously [18]. This would require more effort than spoofing a single trait, which could be a valid deterrent for discouraging the adversary from attempting the attack. However, the above belief has been questioned very recently by the results in [29], and subsequently in [19, 29], that showed that multimodal systems can be evaded by spoofing *only one* biometric trait. The scope of these results is limited to the "worst-case" scenario in which the attacker is able to fabricate a "perfect" replica of the trait of the targeted client, namely, a fake trait whose matching score follows the same distribution of that of genuine users. If these results turned out to hold in general, this would imply that multimodal systems are actually just a deterrent rather than a real defense against spoofing attacks. It is thus of great interest to investigate whether the above results hold also under realistic scenarios, where the worst-case assumption above does not necessarily hold.

To improve the robustness of multimodal systems to spoofing attacks, two novel score fusion rules were proposed in [29]. Basically, contrary to state-of-the-art rules like the ones described in Sect. 3.1, they take into account the possibility that one or more sensors are subject to a spoofing attack. One of these rules is of particular interest, since

it is a modification of the LLR rule, and will be used in our experiments:

Extended LLR (ExtLLR). The underlying idea of ExtLLR is to explicitly take into account the presence of fake traits when modeling the impostor score distribution. To this end, the following model was proposed in [29].[2] Let the random variable $U \in \{\mathrm{G}, \mathrm{I}\}$ denote whether a user is a genuine or an impostor. Given a multimodal system made up of M matchers, their scores are assumed to be conditionally independent, given U: $p(s_1, s_2, \ldots, s_M|U) = p(s_1|U)p(s_2|U)\cdots p(s_M|U)$. M binary random variables T_i are then introduced, to denote whether a user is attempting a spoofing attack against the i-th matcher ($T_i = 1$), or not ($T_i = 0$). Genuine users are assumed to always submit a real biometric trait, namely: $P(T_1 = 0, \ldots, T_M = 0|U = \mathrm{G}) = 1$. Furthermore, it is assumed that each of the $2^M - 1$ possible combinations of attacks against one or more matchers are equiprobable. Denoting with α the prior probability of a spoofing attack, this implies:

$$P(T_1, \ldots, T_M|U = \mathrm{I}) = \begin{cases} 1 - \alpha & \text{if } T_i = 0, i = 1, \ldots, M, \\ \frac{\alpha}{2^M - 1} & \text{otherwise.} \end{cases} \tag{6}$$

Finally, M binary random variables F_i are further introduced, to denote whether a spoofing attack carried out by an impostor against the i-th matcher (i.e., when $U = \mathrm{I}$ and $T_i = 1$) is "successful" ($F_i = 1$) or not ($F_i = 0$), in the sense defined below. Clearly, $F_i = 0$ when the i-th matcher is not under attack, which implies $P(F_i = 0|T_i = 0) = 1$. The probability of success $P(F_i = 1|T_i = 1)$ is denoted with c_i. Its value was related to the "security" ot the corresponding matcher. In [29] it is pointed out that evaluating the security of a matcher is a very difficult problem, if not impossible to solve, and thus c_i must be evaluated based on general knowledge about the biometrics at hand. In the experiments of [29], such value was manually set (see below).

It is now possible to derive the conditional distribution of scores that is needed by the standard LLR rule (see Eq. 5), by marginalising over the $2M$ random variables T_i and F_i:

$$\begin{aligned} p(s_1, \ldots, s_M|U) &= \textstyle\sum_{T_1,\ldots,T_M} \sum_{F_1,\ldots,F_M} p(s_1, \ldots, s_M, T_1, \ldots, T_M, F_1, \ldots, F_M|U) \\ &= \textstyle\sum_{T_1,\ldots,T_M} \sum_{F_1,\ldots,F_M} P(T_1, \ldots, T_M|U) \times \prod_{i=1}^{M} \left[P(F_i|T_i)P(s_i|F_i, U)\right]. \end{aligned} \tag{7}$$

To evaluate the above probability, it is necessary to know the M distributions $P(s_i|F_i, U)$. Given the above assumptions, for genuine users ($U = \mathrm{G}$) we have $F_i = 0$, and thus $P(s_i|F_i = 0, U = \mathrm{G})$ can be learnt from genuine training samples, as in the standard LLR rule. For impostor users ($U = \mathrm{I}$) two assumptions are made in [29]. First, in the case of unsuccessful attacks, the conditional score distribution $P(s_i|F_i = 0, U = \mathrm{I})$ is identical to the one of impostors users that do not attempt spoofing attacks, also called "zero-effort" impostors in [19]. Therefore, this distribution can be learnt from traning data as well. Second, the score distribution of successful spoofing attacks is identical to the one of genuine scores: $P(s_i|F_i = 1, U = \mathrm{I}) = P(s_i|F_i = 0, U = \mathrm{G})$. The latter assumption corresponds to the

[2] The availability of a quality score for each matcher was considered in [29], together with the matching score. In the description of the ExtLLR rule we omit the quality score, since it was not used in our experiments, for the reasons explained in Sect. 4.

"worst-case" scenario mentioned above.

It immediately follows that, for a bimodal system ($M = 2$) as the one considered in [29] and in this work, the expression of the joint likelihood in (7) is:

$$
\begin{aligned}
p(s_1, s_2|I) &= \frac{\alpha}{3}(1-c_1)(1+c_2)p(s_1|G)p(s_2|I) && (8)\\
&+ \frac{\alpha}{3}(1+c_1)(1-c_2)p(s_1|I)p(s_2|G) && (9)\\
&+ \frac{\alpha}{3}(1-c_1)(1-c_2)p(s_1|G)p(s_2|G) && (10)\\
&+ \frac{\alpha}{3}(c_1+c_2+c_1c_2)p(s_1|I)p(s_2|I) && (11)\\
&+ (1-\alpha)p(s_1|I)p(s_2|I)\ . && (12)
\end{aligned}
$$

In particular, terms (8) and (9) are related to successful spoofing attempts against a single trait (respectively, trait 1 and 2), (10) corresponds to a successful spoofing attempt against both traits, (11) accounts for unsuccessful spoof attempts against both traits, and (12) corresponds to zero-effort impostor attempts.

In the experiments of [29] a bimodal system made up of a face and a fingerprint matcher was considered (as in this paper). The face matcher was deemed less secure than the fingerprint matcher. To encode this assumption, the corresponding values of the c_i parameters were manually set respectively to 0.3 and 0.7. It was also pointed out that the prior probability of a spoof attack, α, is application dependent, and can be even variable over time in a given application. For the purpose of the experiments in [29], the value of α was set to 0.01. In general, the prior probability of spoofing attacks should be evaluated by the designer of a biometric system, taking also into account the desired level of security. In other words, the higher the α value used in the ExtLLR rule, the lower should be the probability that an impostor user attempting a spoof attack is accepted as genuine. However, this can also increase the FRR of the system.

A different approach was proposed in [19] to improve the robustness of *any* score fusion rule. It consists of setting the operating point of the system, namely, the value of the decision threshold s^* on the fused score $f(s_1, s_2)$ (see Sect. 3.1), to attain a given trade-off between the false rejection rate (FRR) and the so-called "spoof false acceptance rate" (SFAR) [19]. The SFAR is the conditional probability that an impostor *attempting a spoofing attack* is wrongly accepted as a client. For instance, if the "equal error rate" (EER) operating point, defined as the point where the FRR equals the false acceptance rate (FAR), should be chosen according to application requirements, the alternative choice suggested in [19] to improve robustness is to choose the point where the FRR equals the SFAR. Similar choices can be made for other application requirements. In practice, this allows one to improve robustness against spoofing attacks (namely, reducing the SFAR), at the expense of a higher FRR.

4 Experiments

In this section, we present our experimental analysis, whose main goal is to verify if the "worst-case" hypothesis made in [29, 19, 28] holds for real spoofing attacks, and if it can be reliably exploited to evaluate the robustness of score fusion rules, and for designing robust ones. Some interesting insights on the development of robust fusion rules are also highlighted on the basis of the reported results.

This section is organized as follows: the data sets used in our analysis are described in Sect. 4.1, the performances of different fusion rules under realistic and worst-case spoofing attacks are reported in Sect. 4.2, and the distributions of real spoofing attacks are eventually shown in Sect. .

4.1 Data sets of spoofed samples

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

Data set	Number of clients	Number of spoof images per client	Number of live images per client
LivDet09-Silicone (catalyst)	142	20	20
LivDet11-Alginate	80	3	5
LivDet11-Gelatin	80	3	5
LivDet11-Silicone	80	3	5
LivDet11-Latex	80	3	5
Photo Attack	40	60	60
Personal Photo Attack	25	3 (avg.)	60
Print Attack	50	12	16

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

Fingerprint spoofing. We extended the experimental analysis presented in our previous work [5] by considering more kinds of fingerprint spoofing materials. The data sets are described in the following.

LivDet09. This data consists of 142 distinct clients (by "client", here, we mean a distinct finger, even if it belongs to the same person). For each "live" finger and its corresponding fake replica, twenty different impressions were acquired in two different sessions, separated by about two weeks. Only four fingers were considered in this case: the left and right index and thumb fingers. To create the fake fingerprints, we followed the consensual method described in Sect. 2.1. The mold was produced using plasticine-like materials, while the spoofs were created with liquid silicone (silicone with catalyst) as the casting material. The fingerprint images were acquired using the well-known Biometrika FX2000 and Italdata ET10 optical sensors, which respectively have a resolution of 569 dpi and 500 dpi, and a sensing area of 13.2×25 mm and approximately 30.5×30.5 mm. This data set was also used for assessing the performance of fingerprint liveness detection systems at the First International Competition on Fingerprint Liveness Detection (LivDet09) [23].

LivDet11. This data set includes 80 clients (distinct fingers). As in the previous case, different impressions of

Figure 2: Left: original template image of a fingerprint of our data set. A spoof of the same fingerprint obtained by using latex (middle), and silicone (right).

the live and fake fingers were acquired in two different sessions, separated by about two weeks. However, all the ten fingers were considered here. The fake fingerprints were created as in the previous case, but using the following casting materials, which are commonly adopted for replicating fingerprints: gelatin, silicone, alginate, and latex. As for LivDet09, the fingerprint images were acquired using the Biometrika and Italdata biometric sensors. This data set was also used as a baseline to compare different fingerprint liveness detection algorithms at the Second International Competition on Fingerprint Liveness Detection (LivDet11) [33], where it has been partially used.

Some sample images from the above described data set are shown in Fig. 2, where the average quality of the provided spoofs can be appreciated. This figure shows the original, "live" client image, beside a replica made up of latex, and a replica made up of silicone. As it can be seen, the latex image is very similar to the original one, whilst the second one is characterized by some artifacts. The fake fingerprints used in this work represent the state-of-the-art in fingerprint spoofing, thus providing a reasonable set of realistic scenarios.

Face spoofing. We experimented here on the two data sets used in [5], called the Photo Attack and the Personal Photo Attack data sets, plus a recently published data set called the Print Attack database [4, 6]. They are described in the following.

Photo Attack and Personal Photo Attack. We collected and built two face data sets including the same clients but two different kinds of face spoofing attacks: the *Photo Attack* and the *Personal Photo Attack* data sets. The "live" face images of each client were collected into two sessions, with a time interval of about two weeks between them, under different lighting conditions and facial expressions.

We then created the spoofed face images for the Photo Attack data set using the "photo attack" method described in [6, 34]. It consists of displaying a photo of the targeted client on a laptop screen (or printing it on paper), and then show it to the camera. In particular, the testing "live" face images of the clients were used to this end. 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 a set of personal photos voluntarily provided by 25 of the 50 clients in our data set. On average, we were able to collect 5 photos for each client. These photos were taken in different times and under different environmental conditions than those of the live templates. This simulates

Figure 3: Left: original template image of one of the users of our live face data set. Middle: spoofed face of the Photo Attack data set, obtained by a photo attack. Right: spoofed face of the Personal Photo Attack data set, obtained by a personal photo voluntarily provided by the same user.

a scenario where the attacker may be able to collect a photo of the targeted client from the Web; for instance, from a social network or from an image search engine.

Fig. 3 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 are often significantly different from the template images of a biometric system. On the other hand, the latter case may be more realistic, as it would be probably easier for an attacker to obtain a photo of the targeted client from the Web, than an image similar to the his template. According to the above observations, we expect that the fake score distribution of our Photo Attack data set (provided by some matching algorithm) will be very similar to that of the genuine users (as verified in Sect. 4.3), whilst the effectiveness of a spoof attack based on personal photos will strongly depend on the ability of the attacker to obtain images similar to the templates used by the system.

Print Attack. After the Competition on Countermeasures to 2D Facial Spoofing Attacks, held in conjunction with the International Joint Conference on Biometrics, in 2011, the Print Attack database was made publicly available [4, 6]. It consists of 200 video clips of printed-photo attack attempts to 50 clients, under different lighting conditions, and of 200 real-access attempts from the same clients. As we need to operate on images, we extracted the "live" and spoofed face images from the corresponding videos. In particular, for each client, we extracted 12 "live" face images and 16 spoofed face images from each video clip, as summarized in Table 1.

4.2 Multimodal systems under spoofing attack

In this section, we conducted a set of experiments to verify whether and to what extent the worst-case assumption for simulating the fake score distributions in [29, 19] holds, in real spoofing attacks. Therefore, we used a similar experimental protocol as in [29, 19], described in the following.

- Due to the absence of multimodal data sets including spoofing attacks, we built $5 \times 3 = 15$ *chimerical* data sets, by randomly associating face and fingerprint images of pairs of clients of the available five fingerprint and three

face data sets. Note that building chimerical data sets is a widely used approach in experimental investigations on multimodal biometrics [30].

- To carry out more runs of the experiments, each chimerical data set was randomly subdivided into five pairs of training and testing sets. Each training set included 40% of the "virtual" clients,[3] while the remaining 60% were used to build the testing set. Furthermore, all the above procedure was repeated five times, for different random associations of face and fingerprint images of pairs of clients (namely, creating different "virtual" 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 testing set performance, over the resulting twenty-five runs.

- The fake matching scores were computed by comparing each fake image of a given client with the corresponding template image.

- We normalized all matching scores in $[0, 1]$ using the min-max technique [30], estimating the normalization parameters on the training set.

- The performance was assessed by computing DET curves (FRR vs FAR). Note that, in the evaluation of spoofing attacks, the DET curve reports FRR vs SFAR, since only non-zero-effort impostors are considered [19]. In both cases, performance increases as the curve gets closer to the origin.

The NIST Bozorth3[4] and the VeryFinger[5] matching algorithms were used for fingerprint verification. They are both based on matching the fingerprint minute details, called "minutiae". However, as they exhibited very similar behaviors, we will only report the results for Bozorth3. The Elastic Bunch Graph Matching (EBGM) algorithm was used for face verification.[6] It is based on representing a face with a graph, whose nodes are the so-called face "landmarks" (centered on the nose, eyes, and other points detected on the face). These nodes are labelled by a feature vector, and are connected by edges representing geometrical relationships among them. We also carried out some preliminary experiments using the Principal Component Analysis (PCA) and the Linear Discriminant Analysis (LDA), which yield again very similar results to that of the EBGM algorithm, and are thus omitted in this paper.

We investigated three attack scenarios using real fake traits: (a) only fingerprints are spoofed, (b) only faces are spoofed, (c) both fingerprints and faces are spoofed (bimodal or double spoofing). For the scenarios (a) and (b), we also evaluated the corresponding worst-case attacks as defined in [29, 28, 19]. Accordingly, fictitious fake scores were generated by randomly drawing a set of genuine matching scores from the testing set.

We considered the fusion rules described in the previous section, as they provide a well representative set of the state-of-the-art in fusion rules: sum, product, weighted sum (LDA), LLR, and Extended LLR. Since the bimodal

[3] The clients of a chimerical data set are usually referred to as "virtual" clients, since they do not correspond to a real person or identity. They are indeed created by randomly associating the biometric traits of different "real" clients.

[4] http://www.nist.gov/itl/iad/ig/nbis.cfm

[5] http://www.neurotechnology.com/verifinger.html

[6] http://www.cs.colostate.edu/evalfacerec/algorithms5.php

system considered in this paper is the same as in [29], we used for the Extended LLR rule the same values of the parameters c_i, i.e., for the probability that a spoofing attack against either matcher is successful (in the sense defined in Sect. 3.2). We also considered the same value of 0.01 as in [29] for the prior probability of a spoofing attack. As explained in Sect. 3.2, the Extended LLR can also take into account a quality score provided by a matcher, if any. However, to evaluate the contribution of the Extended LLR rule to the robustness of the LLR rule, due only on its capability to model the presence of spoofed samples, in our experiments we did not consider quality scores.

For the sake of simplicity, and without loosing generality, we only report here a representative set of results. In particular, we only report the results for the chimerical data sets obtained by combining the following fingerprint and face data sets (i.e., a subset of all possible 15 combinations):

1. LivDet11-Latex and Photo Attack (Fig. 5, first row, and Table 2);
2. LivDet11-Gelatin and Print Attack (Fig. 5, second row, and Table 3);
3. LivDet11-Silicone and Print Attack (Fig. 5, third row, and Table 4);
4. LivDet11-Alginate and Personal Photo Attack (Fig. 5, fourth row, and Table 5);
5. LivDet09-Silicone (catalyst) and Personal Photo Attack (Fig. 5, fifth row, and Table 6).

The DET curves attained by the considered fusion rules on each of the above data sets are reported in Fig. 5. However, for the sake of space, we did not report the DET curves for the sum rule, as it performed poorly on the considered data sets, contrary to the results in [19]. The reason is that, for any of the considered data sets, the performance of the face matcher was considerably worse than that of the fingerprint matcher. This performance imbalance strongly affected the performance of the sum rule, but not that of the product rule (although one may think that the product rule should be similarly affected), as exemplified in Fig. 4. In fact, the (hyperbolic) decision functions provided by the product rule correctly assigned a very low matching score to the majority of impostors, biased by the very low output of the fingerprint matcher. Conversely, on average, the sum rule increased their score, worsening the performance.

Additionally, in Tables 2-6, we report the performance attained on each data set by all fusion rules, including the sum rule, for different operating points (i.e., decision thresholds). This allows us to compare more directly performance (in terms of FAR and FRR) and robustness to spoofing attacks (in terms of SFAR) of the different fusion rules, besides making the results better accessible. Furthermore, the tables also give information about the standard deviation of FRR, FAR and SFAR, which is not provided by the DET curves. We considered the following three operating points: EER (when FAR=FRR), FAR=1%, FAR=0.1%. Each operating point was fixed on the DET curve obtained without spoofing attacks, namely, the one attained by considering genuine users and zero-effort impostors. The FRR at each selected operating point is reported in the first column of Tables 2-6 (*no spoof*). Then, we computed the SFAR attained

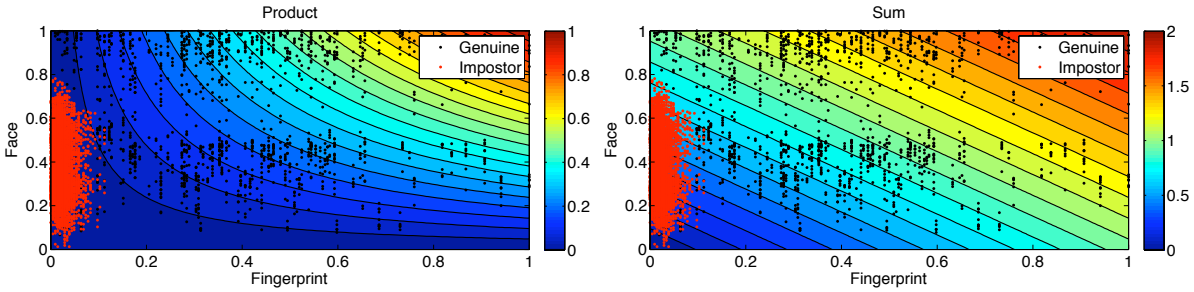

Figure 4: Fusion of face and fingerprint matching scores through product and sum. The values attained by the two fusion rules are shown in different colors. Genuine and impostor scores for Livdet11 (fingerprint) and Print Attack (face) are also reported to highlight how the product rule may outperform the sum rule.

by the different spoofing attacks at the same operating point (reported in the remaining columns). This indeed provides a complete understanding of performance and robustness of each fusion rule: once the operating point is fixed, the effect of spoofing is only to increase the FAR (actually, the SFAR) as it only affects impostor matching scores, while the FRR remains constant.

Rule	*no spoof*	*face*	*w-face*	*fing.*	*w-fing.*	*both*
	EER %	SFAR %	SFAR %	SFAR %	SFAR %	SFAR %
Sum	9.98 ± 2.1	33.25 ± 3.9	37.82 ± 3.8	44.07 ± 4.8	79.85 ± 3.8	60.89 ± 2.9
Product	3.49 ± 1.4	5.72 ± 2.1	6.43 ± 2.2	70.06 ± 5.4	96.11 ± 1.8	73.10 ± 4.9
LDA	3.32 ± 1.5	8.39 ± 4.3	9.87 ± 4.8	70.79 ± 5.6	96.36 ± 2.2	74.09 ± 5.4
LLR	3.60 ± 1.4	5.58 ± 2.8	6.36 ± 3.2	71.41 ± 5.1	96.46 ± 2.2	73.47 ± 5.1
Ext. LLR	3.61 ± 1.4	5.64 ± 2.7	6.40 ± 3.1	71.49 ± 5.0	96.38 ± 2.2	73.57 ± 5.1
	FRR % at FAR=1%	SFAR %	SFAR %	SFAR %	SFAR %	SFAR %
Sum	17.41 ± 3.2	15.58 ± 1.6	20.46 ± 1.9	28.38 ± 5.2	69.00 ± 6.1	46.00 ± 3.8
Product	5.15 ± 2.7	1.93 ± 0.4	2.28 ± 0.5	63.22 ± 4.7	94.37 ± 3.1	66.57 ± 4.4
LDA	5.05 ± 2.6	2.17 ± 0.5	2.73 ± 0.7	64.91 ± 4.7	95.12 ± 3.1	67.83 ± 4.6
LLR	5.46 ± 2.6	1.22 ± 0.4	1.43 ± 0.5	64.94 ± 4.7	95.22 ± 3.1	66.38 ± 4.7
Ext. LLR	5.63 ± 2.8	1.17 ± 0.4	1.38 ± 0.5	64.68 ± 4.8	94.94 ± 3.3	66.03 ± 4.8
	FRR % at FAR=0.1%	SFAR %	SFAR %	SFAR %	SFAR %	SFAR %
Sum	22.76 ± 3.5	8.84 ± 1.1	12.68 ± 1.5	21.65 ± 4.9	62.68 ± 6.7	37.30 ± 3.9
Product	8.59 ± 4.1	0.30 ± 0.1	0.36 ± 0.1	53.41 ± 5.5	90.62 ± 4.1	56.79 ± 4.9
LDA	7.99 ± 3.7	0.26 ± 0.1	0.32 ± 0.2	56.32 ± 5.3	92.45 ± 3.8	58.66 ± 5.2
LLR	8.91 ± 4.1	0.14 ± 0.1	0.17 ± 0.1	56.23 ± 6.0	92.39 ± 3.9	57.48 ± 6.0
Ext. LLR	9.46 ± 5.2	0.16 ± 0.1	0.19 ± 0.1	56.13 ± 5.7	90.97 ± 5.5	57.27 ± 6.0

Table 2: EER, FRR at FAR=1%, and FRR at FAR=0.1% for the considered fusion rules on Livdet11-Latex and Photo Attack (*no spoof*). The SFAR corresponding to the same operating points is reported for real spoofing of fingerprint (*fing.*), face (*face*), and both traits (*both*), and under simulated worst-case spoofing of fingerprint (*w-fing.*), and face (*w-face*). Results are averaged over 25 runs and reported as mean and standard deviation.

In the following, we discuss our results. First, we point out in which scenarios the worst-case assumption provides a good approximation of the performance of multimodal systems under spoofing attacks. Second, we analyze performance and robustness to spoofing of the different fusion rules.

Rule	*no spoof*	*face*	*w-face*	*fing.*	*w-fing.*	*both*
	EER %	SFAR %	SFAR %	SFAR %	SFAR %	SFAR %
Sum	14.35 ± 2.2	46.31 ± 3.0	47.26 ± 2.6	29.98 ± 3.0	76.38 ± 5.1	58.97 ± 3.9
Product	5.25 ± 1.6	16.31 ± 4.4	21.20 ± 5.5	53.54 ± 8.6	92.94 ± 3.0	68.01 ± 7.9
LDA	4.32 ± 1.8	29.54 ± 9.9	38.27 ± 8.5	53.28 ± 10.3	94.02 ± 3.4	69.51 ± 10.1
LLR	4.16 ± 1.6	17.68 ± 8.7	28.65 ± 12.3	56.31 ± 9.5	94.88 ± 2.7	66.64 ± 10.8
Ext. LLR	4.18 ± 1.6	16.52 ± 7.9	27.68 ± 12.3	56.02 ± 9.6	94.84 ± 2.8	66.16 ± 10.9
	FRR % at FAR=1%	SFAR %	SFAR %	SFAR %	SFAR %	SFAR %
Sum	28.04 ± 5.1	32.84 ± 2.3	39.50 ± 1.0	8.42 ± 2.8	56.24 ± 8.9	40.90 ± 3.5
Product	8.87 ± 3.2	4.87 ± 0.9	7.50 ± 1.5	38.40 ± 7.7	88.61 ± 4.9	52.89 ± 6.7
LDA	6.43 ± 2.9	10.42 ± 4.1	23.48 ± 7.8	41.88 ± 7.7	91.61 ± 4.7	56.57 ± 6.5
LLR	6.58 ± 3.0	3.34 ± 1.4	8.18 ± 5.9	45.46 ± 7.9	92.98 ± 4.1	52.52 ± 8.2
Ext. LLR	6.64 ± 3.0	3.15 ± 1.3	7.03 ± 4.8	45.39 ± 7.8	92.95 ± 4.2	52.22 ± 8.0
	FRR % at FAR=0.1%	SFAR %	SFAR %	SFAR %	SFAR %	SFAR %
Sum	34.52 ± 5.7	22.08 ± 4.5	36.08 ± 1.4	3.86 ± 1.9	46.04 ± 9.8	31.93 ± 4.6
Product	13.82 ± 4.2	1.05 ± 0.3	1.90 ± 0.4	26.08 ± 7.2	82.45 ± 6.5	38.21 ± 7.2
LDA	9.98 ± 3.7	1.39 ± 0.6	4.61 ± 2.9	31.64 ± 7.8	88.62 ± 5.6	41.27 ± 7.0
LLR	10.73 ± 4.3	0.37 ± 0.2	0.81 ± 0.5	33.78 ± 8.2	89.53 ± 5.3	38.04 ± 8.4
Ext. LLR	10.61 ± 4.4	0.40 ± 0.2	0.81 ± 0.4	34.15 ± 7.9	89.67 ± 5.4	38.59 ± 7.9

Table 3: EER, FRR at FAR=1%, and FRR at FAR=0.1% for the considered fusion rules on Livdet11-Gelatin and Print Attack (*no spoof*). The SFAR corresponding to the same operating points is reported for real spoofing of fingerprint (*fing.*), face (*face*), and both traits (*both*), and under simulated worst-case spoofing of fingerprint (*w-fing.*), and face (*w-face*). Results are averaged over 25 runs and reported as mean and standard deviation.

Rule	*no spoof*	*face*	*w-face*	*fing.*	*w-fing.*	*both*
	EER %	SFAR %	SFAR %	SFAR %	SFAR %	SFAR %
Sum	13.97 ± 2.0	46.60 ± 1.7	47.39 ± 1.5	29.38 ± 3.1	75.79 ± 5.3	58.42 ± 2.8
Product	4.54 ± 1.2	14.86 ± 3.2	19.35 ± 4.0	46.83 ± 4.7	92.85 ± 2.7	56.96 ± 5.1
LDA	3.77 ± 1.3	27.11 ± 5.7	37.13 ± 4.3	48.60 ± 5.1	94.19 ± 2.7	63.01 ± 4.5
LLR	3.74 ± 1.3	15.03 ± 6.1	27.44 ± 7.7	50.15 ± 5.2	94.99 ± 2.5	57.66 ± 6.2
Ext. LLR	3.75 ± 1.3	14.87 ± 5.8	26.64 ± 8.7	50.17 ± 5.0	94.98 ± 2.4	57.66 ± 5.5
	FRR % at FAR=1%	SFAR %	SFAR %	SFAR %	SFAR %	SFAR %
Sum	27.64 ± 4.7	33.05 ± 1.9	39.29 ± 0.9	8.83 ± 2.5	56.17 ± 8.7	41.14 ± 2.9
Product	7.87 ± 2.5	4.95 ± 1.0	7.64 ± 1.9	37.67 ± 5.0	88.10 ± 4.3	46.66 ± 5.6
LDA	5.70 ± 2.4	10.54 ± 5.2	21.98 ± 9.3	42.78 ± 5.2	91.68 ± 3.9	51.79 ± 6.1
LLR	5.77 ± 2.7	3.53 ± 1.4	8.95 ± 6.4	45.04 ± 5.0	92.95 ± 3.8	48.41 ± 5.5
Ext. LLR	5.80 ± 2.8	3.25 ± 1.2	7.41 ± 5.2	44.93 ± 4.9	92.91 ± 3.8	48.14 ± 5.4
	FRR % at FAR=0.1%	SFAR %	SFAR %	SFAR %	SFAR %	SFAR %
Sum	34.68 ± 5.6	22.22 ± 3.5	35.63 ± 1.4	3.97 ± 1.8	46.04 ± 9.7	31.62 ± 3.6
Product	14.05 ± 3.4	0.98 ± 0.3	1.80 ± 0.6	26.24 ± 4.6	80.65 ± 5.4	36.30 ± 5.2
LDA	9.56 ± 3.4	1.33 ± 0.7	4.69 ± 4.3	34.94 ± 5.1	88.05 ± 4.7	41.01 ± 5.4
LLR	9.71 ± 3.9	0.37 ± 0.1	0.79 ± 0.4	38.95 ± 4.9	89.37 ± 4.8	41.09 ± 5.2
Ext. LLR	9.59 ± 3.9	0.38 ± 0.1	0.77 ± 0.3	39.02 ± 5.0	89.47 ± 4.7	41.22 ± 5.2

Table 4: EER, FRR at FAR=1%, and FRR at FAR=0.1% for the considered fusion rules on Livdet11-Silicone and Print Attack (*no spoof*). The SFAR corresponding to the same operating points is reported for real spoofing of fingerprint (*fing.*), face (*face*), and both traits (*both*), and under simulated worst-case spoofing of fingerprint (*w-fing.*), and face (*w-face*). Results are averaged over 25 runs and reported as mean and standard deviation.

Rule	*no spoof*	*face*	*w-face*	*fing.*	*w-fing.*	*both*
	EER %	SFAR %	SFAR %	SFAR %	SFAR %	SFAR %
Sum	10.57 ± 1.5	10.75 ± 3.5	37.97 ± 4.1	14.80 ± 1.6	78.32 ± 3.2	16.63 ± 4.0
Product	4.08 ± 1.1	6.12 ± 1.8	8.05 ± 2.0	25.09 ± 6.0	95.62 ± 1.6	30.36 ± 9.0
LDA	3.89 ± 1.3	6.48 ± 2.4	11.63 ± 3.7	25.16 ± 5.6	95.81 ± 1.8	28.75 ± 7.5
LLR	4.14 ± 1.1	5.03 ± 1.9	7.89 ± 2.4	25.43 ± 6.0	95.97 ± 1.7	28.52 ± 6.9
Ext. LLR	4.14 ± 1.1	5.17 ± 1.7	8.31 ± 2.8	25.88 ± 5.5	96.03 ± 1.7	28.78 ± 7.1
	FRR % at FAR=1%	SFAR %	SFAR %	SFAR %	SFAR %	SFAR %
Sum	18.80 ± 3.0	1.40 ± 1.0	20.37 ± 2.3	2.83 ± 0.8	66.84 ± 4.8	2.94 ± 1.6
Product	6.38 ± 2.2	1.47 ± 0.6	2.33 ± 0.3	13.61 ± 4.4	93.18 ± 2.8	15.36 ± 6.1
LDA	6.21 ± 2.4	1.47 ± 0.7	2.84 ± 0.8	14.53 ± 4.4	94.09 ± 2.8	15.48 ± 6.2
LLR	6.64 ± 2.5	1.15 ± 0.5	1.71 ± 0.4	15.01 ± 4.7	94.41 ± 2.8	14.81 ± 5.9
Ext. LLR	6.63 ± 2.5	1.13 ± 0.5	1.64 ± 0.3	14.86 ± 4.6	94.29 ± 2.9	14.64 ± 5.8
	FRR % at FAR=0.1%	SFAR %	SFAR %	SFAR %	SFAR %	SFAR %
Sum	24.59 ± 3.3	0.17 ± 0.2	12.53 ± 1.4	0.82 ± 0.5	60.14 ± 5.0	0.66 ± 0.7
Product	9.81 ± 3.1	0.17 ± 0.1	0.38 ± 0.1	6.27 ± 3.1	89.14 ± 3.6	6.05 ± 3.9
LDA	9.21 ± 3.1	0.17 ± 0.1	0.36 ± 0.2	7.23 ± 3.2	91.26 ± 3.5	6.40 ± 3.9
LLR	10.55 ± 4.6	0.15 ± 0.1	0.15 ± 0.1	6.82 ± 3.8	90.86 ± 3.9	5.52 ± 3.8
Ext. LLR	10.85 ± 6.5	0.15 ± 0.1	0.18 ± 0.1	7.33 ± 3.8	88.69 ± 10.8	6.24 ± 3.7

Table 5: EER, FRR at FAR=1%, and FRR at FAR=0.1% for the considered fusion rules on Livdet11-Alginate and Personal Photo Attack (*no spoof*). The SFAR corresponding to the same operating points is reported for real spoofing of fingerprint (*fing.*), face (*face*), and both traits (*both*), and under simulated worst-case spoofing of fingerprint (*w-fing.*), and face (*w-face*). Results are averaged over 25 runs and reported as mean and standard deviation.

Rule	*no spoof*	*face*	*w-face*	*fing.*	*w-fing.*	*both*
	EER %	SFAR %	SFAR %	SFAR %	SFAR %	SFAR %
Sum	15.14 ± 3.2	19.80 ± 6.0	43.97 ± 5.2	18.56 ± 2.7	66.45 ± 6.9	21.75 ± 6.9
Product	1.89 ± 0.7	2.55 ± 1.5	3.88 ± 1.3	24.32 ± 5.2	96.82 ± 1.1	25.61 ± 15.1
LDA	1.70 ± 0.7	1.50 ± 0.9	2.31 ± 1.7	22.21 ± 6.3	96.80 ± 1.7	20.91 ± 14.7
LLR	1.78 ± 0.7	1.96 ± 1.2	2.67 ± 1.1	25.57 ± 5.7	97.46 ± 1.0	24.45 ± 14.8
Ext. LLR	1.79 ± 0.7	1.95 ± 1.2	2.60 ± 1.0	25.50 ± 5.6	97.44 ± 1.1	24.43 ± 14.8
	FRR % at FAR=1%	SFAR %	SFAR %	SFAR %	SFAR %	SFAR %
Sum	28.82 ± 6.2	1.68 ± 0.7	21.47 ± 2.5	2.25 ± 0.8	42.40 ± 11.7	2.27 ± 2.0
Product	2.49 ± 1.2	1.25 ± 0.5	2.16 ± 0.3	19.54 ± 4.7	95.95 ± 1.7	20.51 ± 13.6
LDA	2.56 ± 1.3	0.98 ± 0.4	1.47 ± 0.6	19.40 ± 5.2	96.42 ± 1.8	18.52 ± 13.2
LLR	2.29 ± 1.1	1.02 ± 0.4	1.43 ± 0.2	21.32 ± 5.1	96.82 ± 1.6	20.40 ± 14.0
Ext. LLR	2.29 ± 1.1	1.02 ± 0.4	1.43 ± 0.2	21.32 ± 5.1	96.75 ± 1.5	20.36 ± 13.9
	FRR % at FAR=0.1%	SFAR %	SFAR %	SFAR %	SFAR %	SFAR %
Sum	35.35 ± 6.5	0.18 ± 0.2	14.04 ± 1.9	0.55 ± 0.4	33.61 ± 11.6	0.64 ± 1.2
Product	4.69 ± 1.8	0.12 ± 0.1	0.34 ± 0.1	9.25 ± 3.6	92.56 ± 2.7	9.05 ± 8.5
LDA	4.36 ± 1.8	0.10 ± 0.1	0.18 ± 0.1	10.17 ± 3.9	94.18 ± 2.4	9.30 ± 8.9
LLR	4.56 ± 1.9	0.08 ± 0.1	0.13 ± 0.0	10.03 ± 4.5	94.01 ± 2.6	9.17 ± 8.8
Ext. LLR	8.28 ± 7.6	0.10 ± 0.1	0.17 ± 0.0	9.61 ± 3.9	84.95 ± 16.4	9.04 ± 8.4

Table 6: EER, FRR at FAR=1%, and FRR at FAR=0.1% for the considered fusion rules on Livdet09-Silicone (catalyst) and Personal Photo Attack (*no spoof*). The SFAR corresponding to the same operating points is reported for real spoofing of fingerprint (*fing.*), face (*face*), and both traits (*both*), and under simulated worst-case spoofing of fingerprint (*w-fing.*), and face (*w-face*). Results are averaged over 25 runs and reported as mean and standard deviation.

From the plots in the first row of Fig. 5, it is possible to appreciate that the DET curves for real face spoofing and worst-case face spoofing are very close, for any of the considered fusion rules. This is also true for the corresponding values in Table 2 (*face* and *w-face* columns). The worst-case assumption is thus realistic when faces are spoofed through a photo attack, using an image similar to the template. In other words, the corresponding fake score distributions are very similar to those of the genuine users. Hence, in this case, modeling fake score distributions as genuine ones, as proposed in [29, 19], is acceptable. The same does not hold however for latex-based fake fingerprints, although they are the highest quality (and most effective) fake fingerprints obtained in our data sets. As it can be seen from the plots in the first row of Fig. 5, and from the values in Table 2 (*fing.* and *w-fing.* columns), in this case the SFAR is clearly overestimated by the worst-case assumption (being equal the FRR).

A similar behaviour to that described above is shown in the plots in the second and third row of Fig. 5, corresponding to the data sets obtained by combining Livdet11-Gelatin or Livdet11-Silicone and Print Attack. However, in these cases, the spoofed traits turned out to be not as good and effective as in the previous case, resulting in a stronger violation of the worst-case assumption. This can also be noted by comparing *face* and *w-face* columns, and *fing.* and *w-fing.* columns in Tables 3 and 4.

Lastly, in the fourth and fifth row of Fig. 5, and in Tables 5 and 6, we report the results attained using the least effective spoofed traits, namely, fake fingerprints constructed with alginate and liquid silicone (with catalyst), and face images obtained from personal photos. Note indeed how the difference between the SFAR attained under the worst-case assumption and the one obtained from real spoofs is almost always even higher than that shown in Tables 3 and 4, both for spoofed faces and for spoofed fingerprints. In particular, in the case of face spoofing, the performance is very close to the one attained without any spoofing attack, while, in the case of fingerprint spoofing, the performance is considerably far from both the performance attained in the worst-case scenario and that attained without spoofing attacks.

To summarize, while the worst-case assumption may hold in some cases for face spoofing, our results provide evidence that is very difficult to fabricate fake fingerprints whose score distribution is similar to that of genuine users.

Let us now compare the different fusion rules used in these experiments. When no spoofing attack is performed, all fusion rules exhibited almost the same performance, except for the sum rule, which performed worse (see Tables 2-6, *no spoof* column). As previously pointed out, this is due to the strong performance imbalance between the fingerprint and the face matcher. In this case, indeed, the fusion rule should be more biased toward the most accurate matcher, to achieve better performance. This turned out to be true for all rules, except for the sum.

On the other hand, this behaviour made the sum rule less vulnerable to both real and worst-case fingerprint spoofing, as it attained the lowest SFAR (see *fing.* and *w-fing.* columns). For the same reason above, the sum rule exhibited the worst SFAR under face spoofing (see *face* and *w-face* columns). No appreciable performance difference was exhibited by the other rules in the presence of spoofing attacks. The only exceptions were provided by the LDA under

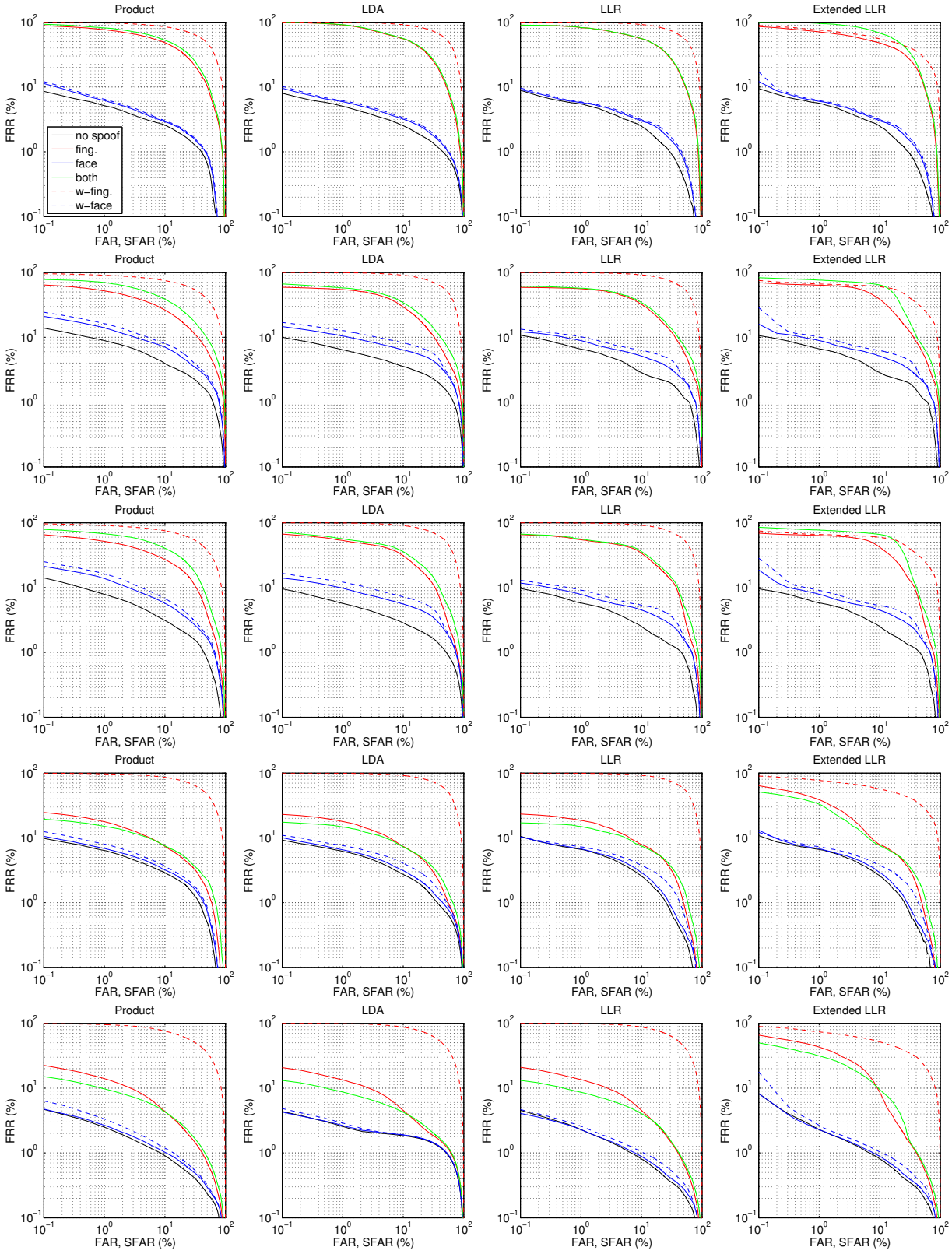

Figure 5: Average DET curves attained on LivDet11-Latex and Photo Attack (first row), LivDet11-Gelatin and Print Attack (second row), LivDet11-Silicone and Print Attack (third row), LivDet11-Alginat and Personal Photo Attack (fourth column), and LivDet09-Silicone (catalyst) and Personal Photo Attack (fifth column). Each column refers to a different fusion rule. Each plot contains the DET curves attained under no spoofing attack (*no spoof*), real spoofing of fingerprint (*fing.*), face (*face*), and both traits (*both*), and under simulated worst-case spoofing of fingerprint (*w-fing.*), and face (*w-face*).

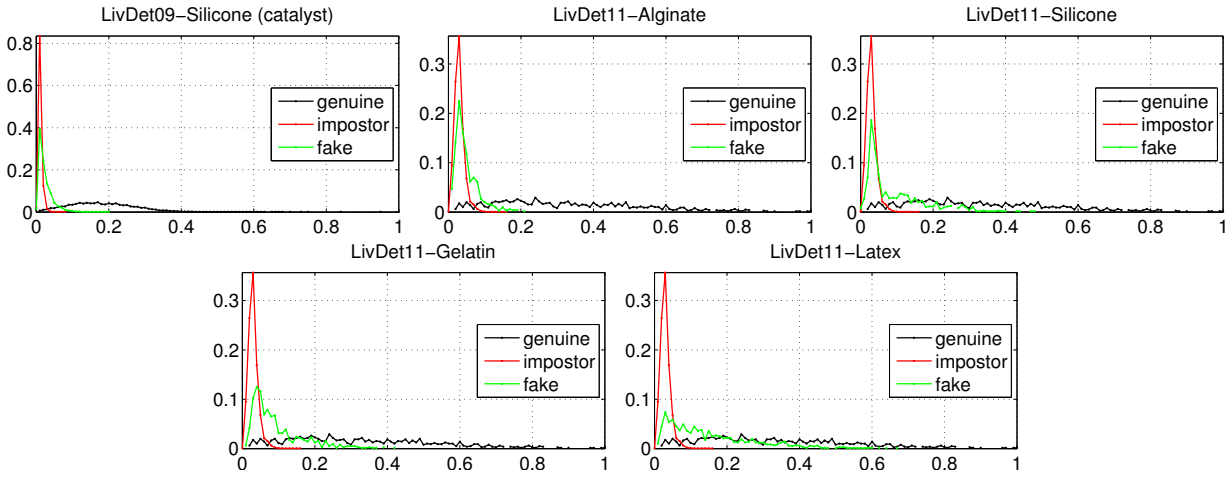

Figure 6: Matching score distributions for the fingerprint data sets, using the Bozorth3 matching algorithm.

real and worst-case spoofing attacks, for Livdet11-Gelatin and Print Attack, and Livdet11-Silicone and Print Attack, at EER and FAR=1% operating points (see Tables 3 and 4, *face* and *w-face* columns).

Surprisingly, for the considered operating points, the Extended LLR performed similarly to the other rules not only in the absence of spoofing attacks, but also in terms of robustness to spoofing, although it was specifically designed to counteract worst-case spoofing attacks (see Tables 2-6). Nevertheless, it is worth noting that such rule even exhibited worse DET curves than the other rules under real spoofing attacks, at very low FAR values; for example, see the case of fingerprint (*fing.*) and double spoofing (*both*) in the plots on the fourth and fifth row of Fig. 5. 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. 3, another problem of the Extended LLR is that setting its parameters (c_1, c_2, and α for a bimodal system) is not trivial, as their values can not be tuned, for instance, on validation data, but can only be hypothesized in advance.

The above results clearly point out that the worst-case assumption is not always suitable for assessing the robustness of multimodal systems to spoofing attacks, as well as to design robust fusion rules against them. They also suggest that a more realistic modeling of the fake score distribution is needed for this purpose.

4.3 Matching score distributions of real spoofing attacks

To further analyse the results reported in the previous section, we report here the matching score distributions of the genuine, impostor, and fake traits, for each fingerprint and face data set, obtained by the Bozorth3 and EBGM matching algorithms, respectively (Figs. 6 and 7). The distributions obtained by the other matching algorithms are very similar, and are not reported for the sake of space.

The worst-case scenario hypothesized in [29, 19] amounts to assuming that the distribution of the fake traits corresponds to that of the genuine users. In the previous section, we already pointed out that this hypothesis can be

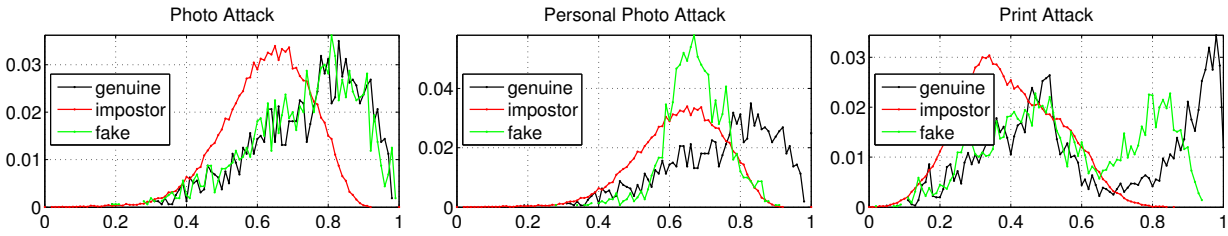

Figure 7: Matching score distributions for the face data sets, using the EBGM matching algorithm.

violated, leading to a too pessimistic evaluation of the SFAR of multimodal biometric systems under spoofing attacks. The matching score distributions in Figs. 6 and 7 confirm the results of the previous section, which are summarized in the following.

(i) The worst-case assumption is too pessimistic, and thus unrealistic, in the case of fingerprint spoofing, even when the fake fingerprints are constructed with the consensual method, as in all our data sets (Fig. 6). The reason is that the image of a fake fingerprint often presents artifacts which affect the matching algorithm; for instance, not all minutiae points can be perfectly replicated from the source image. Nevertheless, the distributions of the fake matching scores may still significantly worsen the performance with respect to the zero-effort impostor distribution, although not to the extent predicted by the worst-case hypothesis in [29, 19]; in particular, this is true when gelatin and latex are used (Fig. 6, second row).

(ii) Conversely, the worst-case assumption is well suited to face spoofing, provided that the fakes are constructed with images that are very similar to the stored templates, as in the case of the Photo Attack and Print Attack data sets (Fig. 7, first and third plot). The reason is that printing a face image on paper, or displaying it on a laptop screen, does not generate any particular artifact which affects the matching algorithm. However, this does not exclude that some particular artifacts may exist (e.g., printing failures or blurring), and, indeed, they can be successfully exploited for liveness detection [4, 34, 6]. This is however not the case, when face images significantly different than the stored templates are used, e.g., when they are collected through the Web, as in the Personal Photo Attack (Fig. 7, second plot).

To summarize, our results confirmed that spoofing attacks against a single biometric trait, either fingerprint or face, may effectively and significantly degrade the performance of a biometric system. However, they also showed that producing very effective fake faces may be relatively easier for an attacker. This is in agreement with the results of the Competition on Countermeasures to 2D Facial Spoofing Attacks [4, 6], and further highlights the need for effective liveness detection techniques against face spoofing. Moreover, this also provides evidence that modelling the matching score distribution of spoofing attacks using the "worst-case" assumption of [29, 19] is not always suitable for evaluating the robustness of multimodal systems, and for developing robust score fusion rules.

5 Conclusions

In this paper, we investigated the robustness of different score fusion rules for multimodal biometric verification systems, against spoofing attacks. In particular, we focused on a bimodal system consisting of fingerprint and face biometrics. A large number of data sets including real spoofing attacks was used for our purpose.

Our results confirmed the conclusion reported in previous works [29, 19, 28], based on *simulated* spoofing attacks, that multimodal systems can be cracked by spoofing a single trait. However, we also provided a clear evidence that the simulated "worst-case" scenario considered in [29, 19, 28] is not always representative of the score distribution of real spoofing attacks. One relevant consequence is that this scenario does not always provide a reliable estimate of the performance drop of a multimodal systems under spoofing attacks. Another consequence is that score fusion rules like the Extended LLR, explicitly designed to deal with spoofing attacks, can be even weaker than standard fusion rules, when the underlying "worst-case" assumption is violated.

We believe that our findings may be exploited both to help system designers and researchers to better evaluate the impact of spoofing attacks, and to develop robust score fusion rules, without the need of actually fabricating spoofed traits. In particular, based on experimental evidences like the ones obtained in this work, more realistic hypothesis on the distribution of the fake traits can be derived, instead of the "worst-case" assumption. This is part of the authors' ongoing work [3, 2].

Acknowledgments

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". The authors would like to thank the anonymous reviewers for their useful comments and suggestions.

References

[1] A. Adler. Vulnerabilities in biometric encryption systems. *5th Int'l Conf. on Audio- and Video-Based Biometric Person Authentication (AVBPA)*, volume 3546 of *LNCS* - Springer, pp. 1100–1109, 2005. 1

[2] Z. Akhtar, G. Fumera, G. L. Marcialis, and F. Roli. Evaluation of multimodal biometric score fusion rules under spoof attacks. In *5th Int'l Conf. on Biometrics (ICB)*. In press, 2012. 23

[3] Z. Akthar, B. Biggio, G. Fumera, and G. L. Marcialis. Robustness of multi-modal biometric systems under realistic spoof attacks against all traits. In *IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications (BioMS)*, pp. 5–10, 2011. 23

[4] A. Anjos and S. Marcel. Counter-measures to photo attacks in face recognition: a public database and a baseline. In *Int'l Joint Conf. on Biometrics (IJCB)*, In press, 2011. 2, 3, 4, 12, 13, 22

[5] B. Biggio, Z. Akthar, G. Fumera, G. L. Marcialis, and F. Roli. Robustness of multi-modal biometric verification systems under realistic spoofing attacks. In *Int'l Joint Conf. on Biometrics (IJCB)*. In press, 2011. 3, 11, 12

[6] M. M. Chakka, A. Anjos, S. Marcel, R. Tronci, D. Muntoni, G. Fadda, M. Pili, N. Sirena, G. Murgia, M. Ristori, F. Roli, J. Yan, D. Yi, Z. Lei, Z. Zhang, S. Z. Li, W. R. Schwartz, A. Rocha, H. Pedrini, J. Lorenzo-Navarro, M. Castrillón-Santana, J. Maatta, A. Hadid, and M. Pietikainen. Competition on counter measures to 2-D facial spoofing attacks. In *Int'l Joint Conf. on Biometrics (IJCB)*, In press, 2011. 2, 3, 4, 5, 6, 12, 13, 22

[7] J. Chirillo and S. Blaul. *Implementing Biometric Security*. Hungry Minds, Incorporated, 1 edition, 2003. 1

[8] S. A. Cole. *Suspect Identities - A History of Fingerprinting and Criminal Identification*. Harvard University Press, 2001. 4, 5

[9] P. Coli, G. L. Marcialis, and F. Roli. Vitality detection from fingerprint images: a critical survey. In *Int'l Conf. on Biometrics (ICB)*, pp. 722–731, 2007. 2, 4

[10] P. Coli, G. L. Marcialis, and F. Roli. Fingerprint silicon replicas: static and dynamic features for vitality detection using an optical capture device. *Int'l J. of Image and Graphics*, 8:495–512, 2008. 2

[11] R. Derakhshani, S. A. C. Schuckers, L. A. Hornak, and L. O. Gorman. Determination of vitality from a non-invasive biomedical measurement for use in fingerprint scanners. *Pattern Recognition*, 36(2):383–396, 2003. 2

[12] J. Galbally, R. Cappelli, A. Lumini, D. Maltoni, and J. Fierrez. Fake fingertip generation from a minutiae template. In *Int'l Conf. on Pattern Recognition (ICPR)*, pp. 1–4, 2008. 4

[13] J. Galbally, C. McCool, J. Fierrez, S. Marcel, and J. Ortega-Garcia. On the vulnerability of face verification systems to hill-climbing attacks. *Pattern Recogn.*, 43(3):1027–1038, 2010. 1

[14] B. Geller, J. Almog, P. Margot, and E. Springer. A chronological review of fingerprint forgery. *J. of Forensic Science*, 44(5):963–968, 1999. 2, 4

[15] A. Godil, S. Ressler, and P. Grother. Face recognition using 3D facial shape and color map information: comparison and combination. In *Biometric Tech. for Human Identification, SPIE*, volume 5404, pp. 351–361, 2005. 5

[16] X. He, Y. Lu, and P. Shi. A fake iris detection method based on fft and quality assessment. In *Chinese Conf. on Pattern Recognition*, pp. 316–319, 2008. 2, 4

[17] A. K. Jain, K. Nandakumar, and A. Nagar. Biometric template security. *EURASIP J. Adv. Signal Process*, 2008:1–17, 2008. 1

[18] A. K. Jain, A. Ross, S. Pankanti, and S. Member. Biometrics: A tool for information security. *IEEE Trans. on Information Forensics and Security*, 1:125–143, 2006. 8

[19] P. Johnson, B. Tan, and S. Schuckers. Multimodal fusion vulnerability to non-zero effort (spoof) imposters. In *IEEE Int'l Workshop on Information Forensics and Security (WIFS)*, pp. 1–5, December 2010. 2, 3, 6, 8, 9, 10, 11, 13, 14, 15, 19, 21, 22, 23

[20] H. Kang, B. Lee, H. Kim, D. Shin, and J. Kim. A study on performance evaluation of the liveness detection for various fingerprint sensor modules. In *7th Int'l Conf. on Knowledge-Based Intelligent Information and Engg. Systems*, pp. 1245–1253, 2003. 2, 4

[21] K. Kollreider, H. Fronthaler, and J. Bigun. Verifying liveness by multiple experts in face biometrics. In *IEEE Computer Vision and Pattern Recognition Workshop on Biometrics*, pp. 1–6, 2008. 2, 6

[22] J. Li, Y. Wang, T. Tan, and A. K. Jain. Live face detection based on the analysis of fourier spectra. In *Biometric Technology for Human Identification, SPIE*, volume 5404, pp. 296–303, 2004. 2, 4

[23] G. L. Marcialis, A. Lewicke, B. Tan, P. Coli, D. Grimberg, A. Congiu, A. Tidu, F. Roli, and S. A. C. Schuckers. First Int'l Fingerprint Liveness Detection Competition - LivDet 2009. In *15th Int'l Conf. Image Analysis and Processing (ICIAP)*, volume 5716 of *LNC* - Springer, pp. 12–23, 2009. 2, 3, 4, 5, 11

[24] M. Martinez-Diaz, J. Fierrez, J. Galbally, and J. Ortega-Garcia. An evaluation of indirect attacks and counter-measures in fingerprint verification systems. *Pattern Recognition Letters*, 32(12):1643 – 1651, 2011. 1

[25] T. Matsumoto, H. Matsumoto, K. Yamada, and S. Hoshino. Impact of artificial "gummy" fingers on fingerprint systems. *Opt. Sec. Counterfeit Deterrence Tech. IV, Proc. SPIE Vol. 4677*, pp. 275–289,, 2002. 4

[26] T. Putte and J. Keuning. Biometrical fingerprint recognition: Don't get your fingers burned. In *4th Working Conf. on Smart Card Research and Advanced Applications*, pp. 289–303, 2000. 4

[27] N. K. Ratha, J. H. Connell, and R. M. Bolle. An analysis of minutiae matching strength. In J. Bigün and F. Smeraldi, editors, *Proc. 3rd Int'l Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA)*, volume 2091 of *LNCS* -Springer, pp. 223–228, 2001. 1, 2

[28] R. N. Rodrigues, N. Kamat, and V. Govindaraju. Evaluation of biometric spoofing in a multimodal system. In *Int'l Conf. Biometrics: Theory Applications and Systems (BTAS)*, pp. 1–5, 2010. 3, 6, 11, 14, 23

[29] R. N. Rodrigues, L. L. Ling, and V. Govindaraju. Robustness of multimodal biometric fusion methods against spoof attacks. *J. of Visual Languages and Computing*, 20(3):169–179, 2009. 2, 3, 6, 8, 9, 10, 11, 13, 14, 15, 19, 21, 22, 23

[30] A. A. Ross, K. Nandakumar, and A. K. Jain. *Handbook of Multibiometrics*. Springer, 2006. 6, 14

[31] L. Thalheim and J. Krissler. Body check: biometric access protection devices and their programs put to the test. *Computer Magazine*, pp. 114–121, 2002. 4

[32] U. Uludag and A. K. Jain. Attacks on biometric systems: A case study in fingerprints. In *Proc. SPIE-EI 2004, Security, Steganography and Watermarking of Multimedia Contents VI*, pp. 622–633, 2004. 1, 2

[33] D. Yambay, L. Ghiani, P. Denti, G. L. Marcialis, F. Roli, and S. Schuckers. LivDet2011 - Fingerprint Liveness Detection Competition 2011. In *5th Int'l Conf. on Biometrics (ICB)*, In press, 2012. 2, 3, 4, 5, 12

[34] Z. Zhang, D. Yi, Z. Lei, and S. Z. Li. Face liveness detection by learning multispectral reflectance distributions. In *Int'l Conf. on Automatic Face and Gesture Recognition*, pp. 436–441, 2011. 2, 5, 6, 12, 22