Robustness Evaluation of Biometric Systems under Spoof Attacks

Zahid Akhtar, Giorgio Fumera, Gian Luca Marcialis, and Fabio Roli

Dept. of Electrical and Electronic Eng., Univ. of Cagliari
Piazza d'Armi, 09123 Cagliari (Italy)
{z.momin,fumera,marcialis,roli}@diee.unica.it
http://prag.diee.unica.it

Abstract. In spite of many advantages, multi-modal biometric recognition systems are vulnerable to spoof attacks, which can decrease their level of security. Thus, it is fundamental to understand and analyse the effects of spoof attacks and propose new methods to design robust systems against them. To this aim, we are developing a method based on *simulating* the fake score distributions of individual matchers, to evaluate the relative robustness of different score fusion rules. We model the score distribution of fake traits by assuming it lies between the one of genuine and impostor scores, and parametrize it by a measure of the relative distance to the latter, named *attack strength*. Different values of the attack strength account for the many different factors which can affect the distribution of fake scores. In this paper we present preliminary results aimed at evaluating the capability of our model to approximate realistic fake score distributions. To this aim we use a data set made up of faces and fingerprints, including realistic spoof attacks traits.

Keywords: Biometric systems, Performance evaluation, Spoof attacks, Adversarial pattern recognition

1 Introduction

Biometrics are biological or behavioural characteristics that are unique for each individual. In order to combat growing security risks in information era, academies, governments and industries have largely encouraged research and adoption of biometric identification systems. The main advantage of biometric technologies compared to conventional identification methods is replacing "what you have" and "what you know" paradigms with "who you are" one, thus preventing identity fraud by using biometrics patterns that are claimed to be hard to forge.

However, several researches have shown that some biometrics, such as face and fingerprint, can be stolen, copied and replicated to attack biometric systems [1, 2]. This attack is known as *spoof attack*, and also named as *direct attack*. It is carried out by presenting replicated biometric trait to the biometric sensor. "Liveness" testing (vitality detection) methods have been suggested among feasible counteractions against spoof attacks. Liveness testing, which aims to detect

whether the submitted biometric trait is live or artificial, is performed by either software module based on signal processing or hardware module embedded into the input device itself [2, 3]. But, so far, the literature review states that no effective method exists yet. Moreover, the collateral effect when biometric systems are coupled with liveness detection methods is the increase of false rejection rate.

In our opinion, it is pivotal to develop also methods, beside liveness detection ones, to design secure biometric systems. A straightforward approach could be to fabricate fake traits to evaluate the security of the system under design. However, constructing reliable fake replicas and simulating all possible ways in which they can be realised, is impractical [1]. A potential alternative is to develop methods based on *simulating* the distribution of fake biometric traits. To the best of our knowledge, no systematic research effort has been carried out toward this direction yet. The only works which addressed this issue are [5, 14, 15], where the fake distribution is simulated by assuming that attacker is able to replicating exactly the targeted biometric (worst-case scenario): in other words, the fake score distribution coincides with that of genuine users.

Based on above motivation, we are currently developing a method for evaluating the robustness of multi-modal systems to spoof attacks, based on simulating the score distribution produced by fake traits at the matchers output, and then on evaluating the relative robustness of different score fusion rules. Due to the unknown impact of several factors, such as particular biometric trait being spoofed, forgery techniques and skills used by the attackers, etc., on position and shape of score distribution, we make substantive assumptions on the potential form and shape it can get. In particular, we argue that the fake score distribution generated by comparing fake replica of a given subject with the corresponding template of that subject, is between impostor and genuine distributions. On the basis of these considerations, as starting point of our research, we model fake scores as a combination of the genuine and impostor ones, on the basis of a single parameter, that we call "attack strength". This parameter controls the degree of similarity of the fake and genuine scores, with respect to the impostor scores. The attack strength quantifies the effect of several factors mentioned above, and allow to figure out more possible scenarios that the only worst-case one [5, 14, 15]. To evaluate the robustness of a given multi-modal system under spoof attacks using our method, the testing impostor scores of the matcher under attack have to be replaced with simulated fake scores generated as mentioned above. The system designer can also evaluate the robustness of the system by repeating the above procedure for different values of attack strength parameter. In this paper, we present preliminary results aimed at evaluating the capability of our model to approximate realistic fake score distributions.

Our model of the fake score distribution is presented in Sect. 2. In Sect. 3 we describe its preliminary experimental validation on two data sets of faces and fingerprints including real spoof attacks.

2 A model of the match score distribution produced by spoof attacks

We denote the output score of a given biometric matcher as random variable s, and denote with G and I the event that the input biometric trait comes respectively from a genuine or an impostor user. The respective score distributions will be denoted as $p(s|\mathrm{G})$ and $p(s|\mathrm{I})$. In the standard design phase of a multi-modal biometric verification systems, the score of the individual matchers $s_1, s_2, \ldots$ are combined using some fusion rule, and a decision threshold t is set on the fused matching score $s_\mathrm{f} = f(s_1, s_2, \ldots)$, so that a user is accepted as genuine if $s_\mathrm{f} \geq t$, and is rejected as an impostor otherwise. The threshold t is usually set according to applications requirements, like a desired false acceptance rate (FAR) or genuine acceptance rate (GAR) value. This defines the so-called operational point of the biometric system. The FAR and GAR values are estimated from training data made up of a set G_tr of genuine scores and a set I_tr of impostor scores.

A straightforward way to analyse the performance of biometric system under spoof attacks is to fabricate fake biometric traits and present them to the system. However, this can be a lengthy and cumbersome task [4]. An alternative solution for multi-modal systems is to *simulate* the effects of spoof attacks on the matching score of the corresponding biometric trait. This is the approach followed in [5, 14, 15]. In these works, the robustness of multi-modal systems against spoof attacks was evaluated in a worst-case scenario, assuming that the matching score produced by a spoofed trait is identical to the score produced by the original trait of the corresponding genuine user. Accordingly, the score distribution of spoofed traits was assumed to be identical to the genuine score distribution.

However, when a fake trait is presented to the biometric sensor, many factors can influence the resulting output score distribution, such as the particular biometric trait spoofed, the forgery approach, the ability of the attacker in providing a "good" biometric trait of the targeted subject as model for his replica, the specific matching algorithm used by the system, the degree of "robustness" of the representation and matcher themselves to noisy patterns, etc. In practice it can be very difficult, if not impossible, to systematically construct fake biometric traits with different degrees of similarity to the original traits. Due to the current very little knowledge on how aforesaid factors affect the fake score distribution, we argue that the only feasible way is to simulate their effect.

A different scenario than the worst-case one considered in [5, 14, 15] could be modelled by considering a score distribution of fake traits lying between the genuine and impostor distributions. For example, in the case of fingerprint, its "similarity" to the impostors distribution will be caused by several factors as artefacts in the replica, the image distortion from the mould to the cast, the good/bad pressure of the attacker on the sensor surface when placing the spoofed fingerprint, whilst its "similarity" to the genuine users one is given by the fact that several important features, as the ridge texture and minutiae locations, will be the same of the correspondent subject. In absence of more specific information on the possible shapes that the fake score distribution may exhibit, we propose

to simulate the one of any individual matcher as follows: denoting the event that the input biometric trait comes from a spoof attack as F, and the corresponding score distribution as $p(s|\mathrm{F})$, we replace each impostor score s_I with a fictitious score s_F given by

$$s_\mathrm{F} = (1-\alpha)s_\mathrm{I} + \alpha s_\mathrm{G} \ , \tag{1}$$

where s_G is a randomly drawn genuine score, and $\alpha \in [0,1]$ is a parameter which controls the degree of similarity of the distribution of fake scores to the one of genuine scores. The resulting distribution of fictitious fake scores $p(s|\mathrm{F})$ is thus "intermediate" between the ones of $p(s|\mathrm{I})$ and $p(s|\mathrm{G})$. By using different values of α, one gets different possible distributions: the higher the α value, the closer $p(s|\mathrm{F})$ to the genuine score distribution $p(s|\mathrm{G})$. Accordingly, we name α "attack strength". This parameter, α, and related Eq. (1), are aimed not to model the physical fake generation process, but only its effect on the corresponding distribution $p(s|\mathrm{F})$, which depends on several causes like the ones mentioned above. In this paper, we want to investigate if the above model allows us to obtain a reasonable approximation of realistic fake score distributions.

Since the designer has no a priori information about the possible characteristics of the attacks the system may be subject to, he should consider several, hypothetical distributions corresponding to different α, and evaluate the robustness of the score fusion rules of interest against each of them.

Accordingly, Algorithm 1 details the proposed procedure. First, the decision threshold t on the combined score s_f has to be estimated from training data made up of a set of genuine and impostor scores, $\mathrm{G_{tr}}$ and $\mathrm{I_{tr}}$, as described above. In the standard performance evaluation procedure, the performance is then evaluated on a distinct test set of genuine and impostor scores, denoted as $\mathrm{G_{ts}}$ and $\mathrm{I_{ts}}$. To evaluate the performance under a spoof attack, we propose instead to replace the impostor scores $\mathrm{I_{ts}}$ corresponding to the matcher under attack with a set of fictitious fake scores $\mathrm{F_{ts}}$, obtained from Eq. (1). This can be done several times, using different α values to evaluate the performance under spoof attacks of different strength.

Note that using Eq. (1), if the randomly chosen genuine score s_G is higher than the impostor score s_I, then the latter is replaced by a greater fictitious fake score s_F. Therefore, for any threshold value t the FAR evaluated under a simulated spoof attack is likely to be higher than the FAR evaluated in the standard way, without spoof attacks. The GAR remains unchanged instead, as spoof attacks do not affect genuine scores. Accordingly, as the value of α increases, the corresponding FAR is likely to increase from the values attained for $\alpha = 0$, corresponding to the absence of attacks, to the worst-case corresponding to $\alpha = 1$. Hence, the above procedure allows one to evaluate how the system's performance degrades for different potential fake score distributions characterised by a different attack strength. In particular, it can be useful to check the amount of the relative "shift" (the corresponding α value) of the impostor score distribution toward the genuine one, such that the system's performance (the FAR) drops below some given value. The more gracefully the performance degrades (namely,

Algorithm 1 Procedure for evaluating the performance of a multi-modal biometric system under a simulated spoof attack

Inputs:

- A training set $(G_{\text{tr}}, I_{\text{tr}})$ and a testing set $(G_{\text{ts}}, I_{\text{ts}})$ made up of N vectors of matching scores coming from genuine and impostor users;
- α: the attack strength value for the matcher under attack.

Output: The system's performance under a simulated spoof attack with attack strength α.

1: Set the threshold t from training data $(G_{\text{tr}}, I_{\text{tr}})$, according to given performance requirements.
2: Replace the scores I_{ts} of the matcher under attack with a same number of fictitious fake scores F_{ts} generated by Eq. (1).
3: Evaluate the performance of the multi-modal system on the scores $(G_{\text{ts}}, F_{\text{ts}})$

the higher the α value for which the FAR drops below some value of interest), the more robust a system is.

3 Experimental Results

In this section, we report a preliminary validation of our model of the fake score distribution of a single matcher, using two data sets including realistic spoof attacks. More precisely, our aim is to investigate whether realistic fake score distributions can be reasonably approximated by our model, for some α values.

3.1 Data sets

Since no biometric data sets including spoof attack samples are available publicly, we collected two sets of face and fingerprint images and created spoof attacks. The collected data set contains face and fingerprint images of 40 individuals, with 40 genuine samples and 40 fake samples (spoof attacks) per individual.

Face images were collected under different facial expressions and illumination. Spoofed face images were created with a "photo attack" [11]: we put in front of the camera the photo of each individual displayed on a laptop screen. For each individual, we created 40 spoofed images.

Fingerprint images were collected using Biometrika FX2000 optical sensor. Fake fingers were created by the consensual method with liquid silicon as carried out in [6–8]. We fabricated fake fingerprint using plasticine-like material as the mould while two-compound mixture of liquid silicon and a catalyst as cast. The main property of the material utilised as the cast is high flexibility silicon resin (SILGUM HF) with a very low linear shrinkage. Further details on fingerprint spoof production can be found in [12].

The fingerprint and the face recognition systems used in the experiments were implemented using the minutiae-based Neurotechnologs VeriFinger 6.0 and the elastic bunch graph matching (EBGM) [9], respectively.

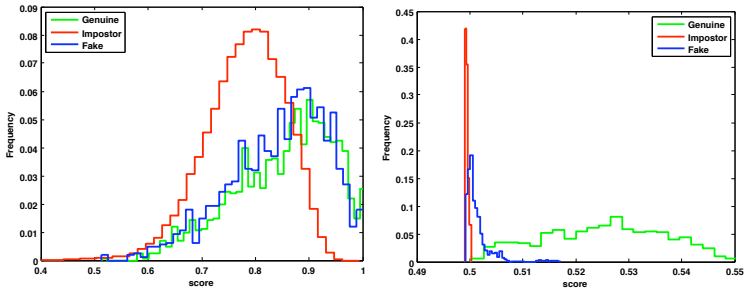

Fig. 1. Histograms of genuine, impostor and fake scores computed with the collected face (left) and fingerprint (right) image data sets.

3.2 Results

In Fig. 1 the histograms of genuine, impostor and fake scores computed with the above data sets are shown. It is worth noting that these distributions exhibit two very different degrees of "attack strength": the fake score distribution of fingerprints is close to the impostor distribution, while the one of faces is much close to the genuine distribution. This provides a first, qualitative support to the assumption behind our model, namely that different, realistic fake score distributions can lie at different relative "distances" from the genuine and impostor ones.

To investigate whether the realistic fake scores distributions of Fig. 1 can be reasonably approximated by our model, for some α value, we evaluated the dissimilarity between them and the ones provided by our model, as a function of the attack strength α, and empirically computed the α value that minimised the dissimilarity between the two distributions. The fictitious fake scores were obtained as described in Algorithm 1. To assess the dissimilarity between the two distributions, we used the L1-norm Hellinger distance [13], also called Class Separation Statistic [10]. The L1-norm Hellinger distance between two probability distribution functions $f(x)$ and $g(x), x \in \mathcal{X}$ can be measured as:

$$\int_{\mathcal{X}} |f(x) - g(x)| \mathrm{d}x.$$

Since this is a non-parametric class separation statistic, it can be used for all possible distributions.

The α values which minimise the dissimilarity between the fake score distribution obtained by our method and the real one is reported in Table 1. The corresponding distributions are depicted in Fig. 2.

Fig. 2 and Table 1 show that our approximation is rather good for the face data set. It is less good for the fingerprint data set instead, but it could be still acceptable to the aim of evaluating the relative robustness of different score fusion rules in a multi-modal system, which is the final aim of this model. Let us better explain this point. Obviously, in practice the designer of a biometric

Data set	Hellinger distance	α
Face	0.0939	0.9144
Fingerprint	0.4397	0.0522

Table 1. Minimum values of the Hellinger distance between the real distribution of fake scores and the one obtained by our model, as a function of α, for the face and fingerprint data sets. The corresponding α value is also shown.

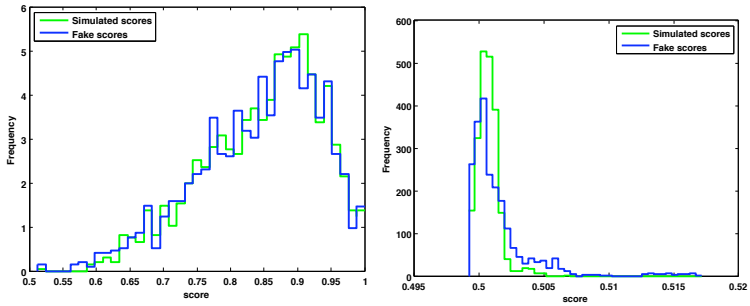

Fig. 2. Probability distributions of the scores of fake faces (left) and of fake fingerprints (right) obtained from our data sets (blue), and obtained by our method for fake score simulation (green), for the α value of Table 1.

system can not not know in advance what shapes the fake score distributions will exhibit, if the system will be subject to a spoof attack. Accordingly, the robustness of a multi-modal system must be evaluated for several α values. What the above results show is a preliminary evidence that the simulated distributions one obtains using our model, for different α values, can actually give reasonable approximations of possible, realistic distributions.

For the sake of completeness, we also evaluated the accuracy of our model of fake score distribution in approximating the performance of the *individual* matcher under attack, for the α values of Table 1 that give the best approximation of the fake score distribution, although this is not the final aim of this model as explained above.

When the system is under a spoof attack, only the False Acceptance Rate (FAR) value changes, while the genuine acceptance rate (GAR) remains unchanged, since it does not depend on the matching scores of the impostors. To check the accuracy of the FAR approximated by our model, we compared the FAR of the mono-modal system attained under real spoof attacks with the FAR provided by our model, for all possible values of the threshold.

Fig. 3 shows the FAR as a function of the threshold for the uni-modal biometric system when no spoof attack is included in the data set (i.e., using only the genuine and impostor data; the "no attack" curve), under a real spoof attack against the face (fingerprint) matcher (using the fake biometric traits of

	Operational point	Real FAR	Approximated FAR (our model)	Approximated FAR (worst-case assumption)
Face System	zeroFAR	0.048	0.042	0.114
	1%FAR	0.235	0.233	0.243
Fingerprint System	zeroFAR	0.506	0.625	0.948
	1%FAR	0.600	0.808	0.951

Table 2. Comparison between the FAR attained at the zeroFAR and 1% FAR operational points by the uni-modal biometric system under a real spoof attack ("real FAR") and the FAR approximated by our model ("approximated FAR").

our data set; the "real spoof attack" curve), and by a simulated spoof attack (using the fake scores provided by our method with the α values of Table 1; the "simulated attack" curve). It can be seen from Fig. 3 that our model provides a quite accurate approximation of the FAR in the case of face spoofing (Fig. 3, left): the maximum absolute difference between the real and the approximated FAR is 0.02. In the case of fingerprint spoofing (Fig. 3, right), our model overestimates the FAR by an amount of up to 0.03 for threshold values lower than 0.502, while it underestimates the FAR up to a larger amount for threshold values greater than 0.502. This is due to the positive skewness of the real fake fingerprint scores, as can bee seen in Fig. 2. Note however that the threshold t corresponding to the zeroFAR operational point is 0.500, as can be seen from Fig. 1 (right). It is worth remarking that zeroFAR is the operational point such that the threshold leads to a zero FAR value on training data, and maximises the correspondent GAR value. Therefore, threshold values more than this one are out of the designer interest and can be neglected. This means that threshold values where the real FAR is underestimated by our model can be neglected as well, since they are localised for threshold values higher than 0.502.

Accordingly, let us focus in particular on high security operational points like the zeroFAR and 1% FAR, which are very crucial in order to assess the system robustness. The corresponding FAR attained by the fake score distribution in our data set ("Real FAR") and the approximated FAR using our model is reported in Table 2. We also report for comparison the approximated FAR obtained using the worst-case assumption of [5, 14, 15]. The reported results show that our method provides a good approximation of the performance of the two considered uni-modal systems under spoof attacks, at these operational points. The overestimation of the values for the fingerprint system is in some sense beneficial, since it puts the designer in the position to expect a performance decrease higher than that occurring in the real case. In addition, it can be seen that our model is more flexible and appropriate for fake score distributions quite far from the worst-case one, as happens for fingerprints.

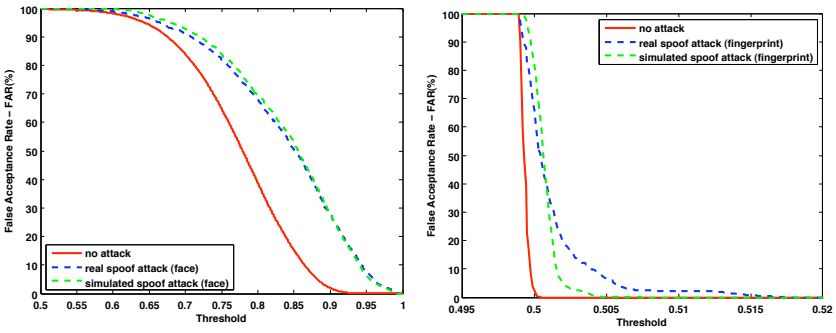

Fig. 3. FAR of the uni-modal biometric systems as a function of the threshold applied to the score, when the data set does not contain spoof attacks ("no attack" curve), under a real spoof attack against the face (left) or fingerprint (right) matcher ("real spoof attack" curve), and under a spoof attack simulated with our method ("simulated attack" curve).

To sum up, our preliminary results provide some evidence that our model is able to reasonably approximate realistic distributions of the matching scores produced by spoof attacks.

4 Conclusions

Assessing the robustness of multi-modal biometric verification systems under spoof attacks is a crucial issue, do to the fact that replicating biometrics is a real menace. The state-of-the-art solves this problem by simulating the effect of a spoof attacks in terms of fake score distribution modelling, for each individual matcher. In particular, the fake score distribution is assumed to be coincident to the genuine users one, thus drawing a worst-case scenario.

However, a more realistic modelling should take into account a larger set of cases. Unfortunately, the approach of fabricating fake biometric traits to evaluate the performance of a biometric system under spoof attacks is impractical. Hence, we are developing a method for evaluating the robustness of multi-modal systems against spoof attacks, based on *simulating* the corresponding score distribution.

In this work we proposed a model of the fake score distribution that accounts for different possible realistic scenarios characterised by factors like different spoofing techniques, resulting in different degrees of similarity between the genuine and the fake score distribution. Such factors are summarised in our model in a single parameter associated to the degree of similarity of the fake score distribution to the genuine one, which is named accordingly "attack strength". A designer may use this method to generate several fake distributions for different α values, to analyse the robustness of the multi-modal system under design.

Preliminary experimental results provided some evidence that our model is capable to give reasonable approximations of realistic fake score distributions,

and also to be a good alternative to the model based on the worst-case scenario adopted so far. Currently, we are working on constructing data sets containing spoofing attacks of different biometric traits, spoofing techniques, matchers, etc., to give a more extensive validation of our model, and to evaluate the effectiveness of our method for robustness evaluation of multi-modal systems under spoof attacks.

Acknowledgment This work was partially supported by the TABULA RASA project, 7th Framework Research Programme of the European Union (EU), grant agreement number: 257289, and 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).

References

1. T. Matsumoto, H. Matsumoto, K. Yamada, S. Hoshino: Impact of artificial "gummy" fingers on fingerprint systems. In: Optical Security and Counterfeit Deterrence Techniques IV, Vol. 4677 of Proc. of SPIE, pp. 275-289 (2002)
2. Y. Kim, J. Na, S. Yoon, J. Yi: Masked Fake Face Detection using Radiance Measurements. J. Opt. Soc. Am. - A, vol. 26, no. 4, pp. 760-766 (2009)
3. H. Kang, B. Lee, H. Kim, D. Shin, J. Kim: A study on performance evaluation of the liveness detection for various fingerprint sensor modules. In: Proc. Seventh Int. Conf. on Know. Based Intel. Info. and Engg. Sys., pp. 1245-1253 (2003)
4. G.L. Marcialis, A. Lewicke, B. Tan, P. Coli, D. Grimberg, A. Congiu, A. Tidu, F. Roli, S. Schuckers: First International Fingerprint Liveness Detection Competition. In: Proc. 14th Intl Conf. on Image Analysis and Proc., pp. 12-23 (2009)
5. R.N. Rodrigues, L.L. Ling, V. Govindaraju: Robustness of Multimodal Biometric Methods against Spoof Attacks. JVLC, vol. 20, no. 3, pp. 169-179 (2009)
6. A. Abhyankar, S. Schuckers: Integrating a Wavelet Based Perspiration Liveness Check with Fingerprint Recognition. Patt. Rec., vol. 42, no. 3, pp. 452-464 (2009)
7. P. Coli, G.L. Marcialis, F. Roli: Fingerprint Silicon Replicas: Static and Dynamic Features for Vitality Detection using an Optical Capture Device. Int'l J. of Image and Graphics, vol. 8, no. 4, pp. 495-512 (2008)
8. G.L. Marcialis, F. Roli, A. Tidu: Analysis of Fingerprint Pores for Vitality Detection. In: Proc. 12th Int'l Conf. on Pattern Recognition, pp. 1289-1292 (2010)
9. D.S. Bolme: Elastic Bunch Graph Matching. Master's Thesis: Dept. of Comp. Science, Colorado State University (2003)
10. A.K. Jain, S. Prabhakar, S. Chen: Combining Multiple Matchers for a High Security Fingerprint Verification System. PRL, vol. 20, no. 11-13, pp. 1371-1379 (1999)
11. G. Pan, Z. Wu, L. Sun: Liveness detection for face recognition. In: Recent Advances in Face Recognition, pp. 236-252 (2008)
12. `http://prag.diee.unica.it/LivDet09`
13. L. LeCam, *Asymptotic Methods in Statistical Decision Theory*, Springer (1986)
14. R.N. Rodrigues, N. Kamat, V. Govindaraju: Evaluation of Biometric Spoofing in a Multimodal System. In: Proc. Fourth IEEE Int. Conf. Biometrics: Theory Applications and Systems, pp. 1-5 (2010)
15. P.A. Johnson, B. Tan, S. Schuckers: Multimodal Fusion Vulnerability to Non-Zero Effort (Spoof) Imposters. In: Proc. IEEE Workshop on Information Forensics and Security, pp. 1-5 (2010)