A. Brömme, N. Damer, M. Gomez-Barrero, K. Raja, C. Rathgeb,
A. Sequeira, M. Todisco, and A. Uhl (Eds.): BIOSIG 2022,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2022

When Facial Recognition Systems become Presentation Attack Detectors

Lázaro J. González-Soler[1], Kevin A. Barhaugen[2], Marta Gomez-Barrero[3], Christoph Busch[1]

Abstract: Recently, biometric systems (BSs) have experienced a broad development mainly due to the great success of deep learning approaches. Generally, most BS provide high security and efficiency. However, they are still vulnerable to attack presentations (APs). To overcome such security issues, these schemes include a Presentation Attack Detection (PAD) module which determines whether the input sample stems from an AP or a bona fide presentation (BP). Traditionally, most PAD subsystems assess the biometric sample prior to the recognition module. In this work, we evaluate to what extent the inverted combination, where the biometric recognition module filters samples prior to the assessment of a PAD mechanism, leads to an overall PAD performance improvement. The experimental evaluation conducted over two well-known databases including challenging attacks, reports a significant improvement in the detection performance when input samples were first filtered by the biometric recognition: only 1% of the APs are accepted while at most 5% BPs are rejected by the PAD subsystem.

Keywords: Biometric systems, Presentation Attack Detection, Face, Interoperability.

1 Introduction

The great success of deep learning in numerous pattern recognition and computer vision tasks has led to the development of high-performance biometric recognition systems. In general, such systems provide high efficiency, security and user convenience. However, they are still vulnerable to attack presentations which have evolved into more sophisticated artefacts or presentation attack instruments (PAIs) over the years. These PAIs can be easily created and launched by a non-authorised subject to gain access to different unattended applications such as financial transactions and smartphone unlocking.

To mitigate such security threats, biometric systems (BSs) are equipped with a presentation attack detection (PAD) module, which aims firstly at determining whether the input sample stems from a live subject (i.e., it is a bona fide presentation - BP) or from an artificial replica (i.e., it is an attack presentation- AP). The pristine facial images, referred to as BPs, are then sent to the biometric recognition (BR) subsystem to verify the claimed identity. This is the traditional combination developed in several applications [CAM13]. However, some studies have shown other configurations where the BR runs in parallel together with

[1] dasec - Biometrics and Internet Security Research Group, Hochschule Darmstadt, Germany, {lazaro-janier.gonzalez-soler;christoph.busch}@h-da.de
[2] NTNU Gjøvik, Norway, kevin.barhaugen@ntnu.no
[3] Hochschule Ansbach, Germany, marta.gomez-barrero@hs-ansbach.de

the PAD scheme [MR14]. In 2012, Marasco *et al.* [MDR12] proposed four different architectures to perform the fusion between fingerprint BR systems and PAD mechanisms. In addition, the authors explored the feasibility of other three methods in terms of biometric performance: *ii*) the BR module is invoked before the PAD method, *iii*) the BR and PAD algorithms run in parallel and the outputs are fused for the final decision, and *iv*) the BR and PAD scores are merged using a Bayesian Belief Network that models causal relationships among them. The latter reported the best performing fusion. Following this idea, Chingovska *et al.* [IAM14, CAM13] evaluated three different fusion techniques to combine the scores produced by the BR and PAD schemes. Other authors have also studied the combination of Multibiometric systems and PAD methods [Ch19], focused mainly on improving the biometric system performance. In addition, they lack a proper evaluation in terms of the metrics defined in the ISO/IEC 30107-3 for biometric PAD [IS17].

Although several studies have been carried out, some questions still remain unanswered: *i*) are BR modules capable of detecting APs? *ii*) can the overall PAD performance benefit from the assessment by the BR? and *iii*) to what extent the BR enhances the PAD performance regarding challenging attacks? To answer these questions, we focus on the two traditional combinations proposed by Marasco *et al.* [MDR12]: *PAD first, BR later* and *BR first, PAD later* (see Fig. 1). The experimental evaluation conducted on two well-known facial databases complies with the ISO/IEC 30107-3 for biometric PAD shows that BRs can be used as AP detectors, improving mainly the detection of challenging attacks.

The remainder of this paper is organised as follows. The experimental setup including the evaluation framework, databases, and metrics is presented in Sect. 2. Sect. 3 discusses the results achieved on different protocols. Finally, conclusions and future work directions are presented in Sect. 4.

2 Experimental Setup

To answer the above questions, we define three goals: *i*) analyse the recognition performance of the selected BR against PAIs, *ii*) establish a benchmark between *PAD first* and *PAD later* settings for known-attacks, and *iii*) for cross-database scenarios.

2.1 Experimental Scheme

To reach our goals, we follow the general overview in Fig. 1 depicting the two schemes assessed in our work. In the *PAD first, BR later* pipeline, the input sample is firstly evaluated by the PAD approach. Those samples classified by the PAD module as BPs are then fed to the BR to verify their identifier. On the other hand, *BR first, PAD later* methodology only determines whether the inputs stem from an AP for those images whose identifier references were previously verified by the BR. In our research, we select as the AP detector the method proposed in [GSGBB21], and ArcFace [De19] is utilised as the BR. These systems have shown remarkable performances for PAD and verification tasks, respectively.

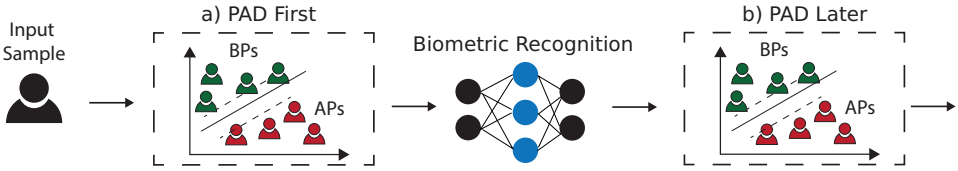


Fig. 1: General overview of the two schemes evaluated in our work: a) *PAD first, BR later* pipeline and b) *BR first, PAD later* pipeline.

Tab. 1: Summary of database characteristics employed in our evaluation.

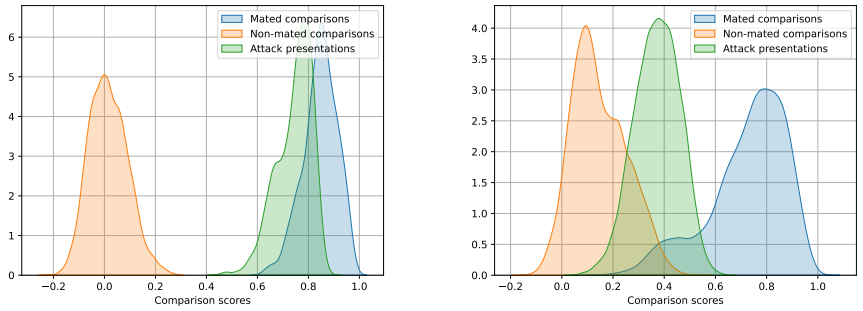
Database	**#Subjects**	**PAI species**	**Train**		**Test**	
			#BPs	#APs	#BPs	#APs
Replay-Mobile	40	Printed, Photo-replay, Video-replay	120	192	110	190
CSMAD-Mobile	8	3D Silicone Mask	178	266	138	364

2.2 Databases

The experimental evaluation is conducted over two freely available databases: Replay-Mobile [Co16] and CSMAD-Mobile [Ra19]. The characteristics of the databases are summarised in Tab. 1. For the PAD evaluation on Replay-Mobile we utilise the "Grandtest" attack protocol described in [Co16]. In addition, we randomly select four subjects for training and the same number for testing over CSMAD-Mobile. In both databases, the training and test sets are created using the same set of PAI species (i.e., it is known-attacks).

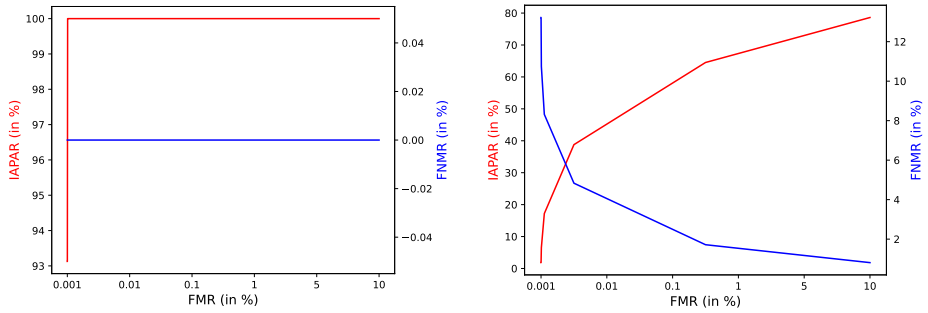
2.3 Metrics

The experimental evaluation is conducted using the metrics defined in the international standard ISO/IEC 30107-3 [IS17] for biometric PAD and ISO/IEC 19795-1 [IS21] for biometric performance: *i*) AP Classification Error Rate (APCER), that is defined as the proportion of APs wrongly classified as BPs; *ii*) BP Classification Error Rate (BPCER), that is the proportion of BPs misclassified as APs; *iii*) False Match Rate (FMR), that is the proportion of the completed biometric non-mated comparison trials that result in a false match; *iv*) False Non-Match Rate (FNMR), which is the proportion of the completed biometric mated comparison trials that result in a false non-match; and *v*) Impostor Attack Presentation Accept Rate (IAPAR) is the proportion of impostor APs using the same PAI species that result in acceptance. Based on these metrics, we report *i*) the Detection Error Trade-off (DET) curves between APCER and BPCER, and between FMR and FNMR; *ii*) the BPCERs observed at different APCER values or security thresholds such as 10% (BPCER10), 5% (BPCER20), and 1% (BPCER100), respectively; and *iii*) the Relative Imposter Attack Presentation Accept Rate (RIAPAR), which is expressed as IAPAR + FNMR for a given threshold. Subsequently, we also report the FNMRs computed at different FMR values such as 0%, 1%, and 10%.



(a) Training set from Replay-Mobile. (b) Training set from CSMAD-Mobile.

Fig. 2: Vulnerability of ArcFace to attack presentations. Best view in a colour version.



(a) IAPAR vs FNMR vs FMR for Replay-Moblie. (b) IAPAR vs FNMR vs FMR for CSMAD-Mobile.

Fig. 3: In-depth analysis of the vulnerability of ArcFace to APs. Best view in a colour version.

3 Results and Discussion

3.1 Biometric System Evaluation

In the first set of experiments, we analyse the biometric performance of ArcFace for non-mated samples and PAIs. For this purpose, we show in Fig. 2-a and b the comparison distributions for mated comparisons (in orange), non-mated comparisons (in blue), and attack attempts (i.e., attack presentations, in green) over the trained subjects. As it can be observed, the non-mated comparisons are generally separated from mated comparisons (orange vs blue, biometric recognition application) and APs (orange vs green). In contrast, the AP distribution overlaps with the BP distribution (blue vs green, PAD application), indicating the need to use an additional PAD method: up to 100% of PAIs could be accepted by the BR depending on the system's threshold.

In addition, the relationship between IAPAR vs. FMR vs. FNMR is represented in Fig. 3-a and b. We observe that the IAPAR increases with the FMR while the FNMR decreases,

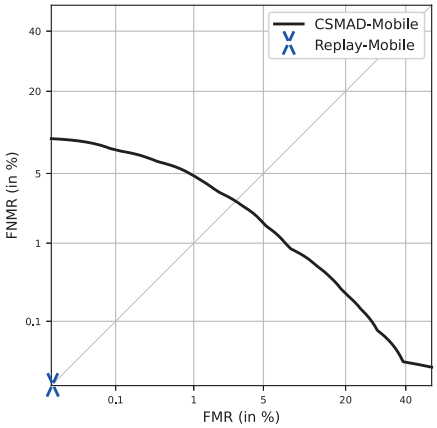


Fig. 4: Recognition performance computed over the CSMAD-Mobile and Replay-Mobile databases. DET curve for Replay-Mobile is not shown, as mated and non-mated comparisons are completed separated.

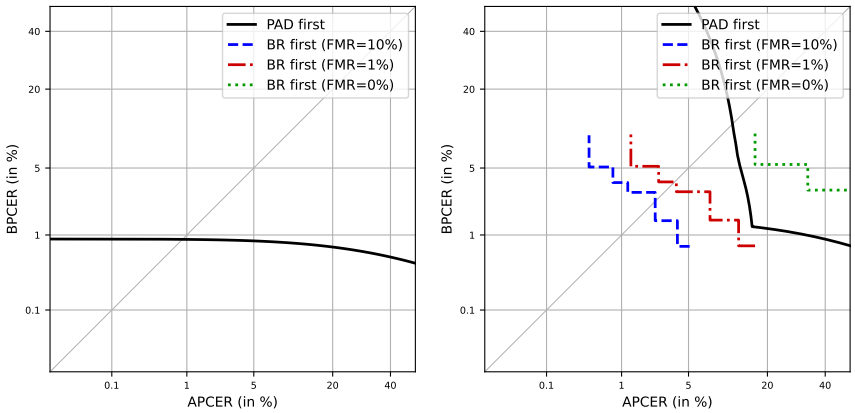
Tab. 2: Characteristics of the test sets after applying *BR first*.

Database	**FMR=10%**		**FMR=1%**		**FMR=0%**	
	#BPs	#APs	#BPs	#APs	#BPs	#APs
Replay-Mobile	110	174	110	174	110	173
CSMAD-Mobile	137	254	135	78	130	6

thereby resulting in RIAPARs of 15.05% and 93.12% for a FMR = 0% on the Replay-Mobile and CSMAD-Mobile, respectively. These results indicate that the selection of a suitable threshold based on the FMR values that minimises the RIAPAR would contribute to improve the detection performance of the PAD module, and thus build a secure and convenient system. It is important to highlight that the curve's behaviour in Fig. 3-a is due to the separability of non-mated comparisons from AP attempts.

3.2 Impact of Biometric Recognition Thresholds

Based on the above observations, we first compute in Fig. 4 the DET curves over the trained samples in the Replay-Mobile and CSMAD-Mobile databases. Since FNMR = 0.0% for any FMR for Replay-Mobile, the DET curve is just a cross at the origin of co-ordinates. In the case of CSMAD-Mobile, we observe that high-security thresholds (i.e., FNMR@FMR $\leq$ 1%) lead to the rejection of up to 13.22% of mated samples. The result difference between Replay-Mobile and CSMAD-Mobile is mainly due to the fact that the samples in the latter were previously preprocessed by the database owners and the face is not always fully visible.

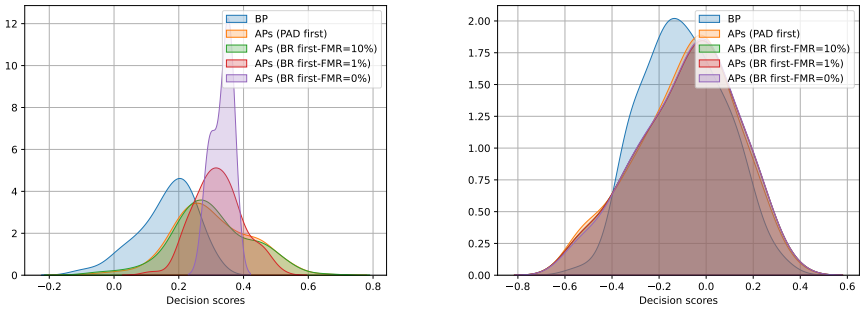


(a) Replay-Mobile. (b) CSMAD-Mobile.

Fig. 5: Detection performance of the PAD algorithm tested on the whole dataset and each filtered set. In Fig. 5-a the DET for *BR first* are missing due to a BPCER = 0.0% for any APCER.

In order to achieve a trade-off between security and user convenience, we select for the next experiments, three different thresholds: FMR=10%, FMR=1%, and FMR=0% and preselect for the PAD evaluation those images in the test set which meet the given threshold. The *BR first, PAD later* methodology is applied in this experiment. In detail, we first select a random BP reference sample per subject in the test set and then compare it with the remaining BPs and APs for the same subject using the traditional cosine metric. Those samples whose cosine value is lower than the input threshold (e.g., FMR=1%) are rejected by the BR and hence removed for the PAD evaluation. We can see in Tab. 2 that the number of BPs and APs filtered by ArcFace is significantly lower than the entire test set (see Tab 1). It may be noted that the BR rejects AP attempts up to 9.0% for Replay-Mobile and 98% for CSMAD-Mobile without sacrificing the user convenience: at most 5.8% of BPs stemming from CSMAD-Mobile are rejected by the module.

Now, we compute the DET curves in Fig. 5 and establish a benchmark of the PAD algorithm tested on the whole set of images (i.e.,*PAD first*, black solid line) with respect to those images filtered by the BR (i.e, *PAD later*, dashed lines). As it may be seen, the PAD algorithm achieves a detection performance improvement for most filtered sets on Replay-Mobile and CSMAD-Mobile. In particular, the evaluated PAD approach yields a BPCER=0.0% for any APCER over the three operating thresholds in Replay-Mobile, thus leading to an accuracy increase regarding the baseline (i.e., *PAD first*). Regarding CSMAD-Mobile, remarkable BPCERs $\leq$ 10.0% for any 0% $\leq$ APCER $\leq$ 10% are obtained for the *PAD later* scheme, which is much lower than the ones yielded over the whole dataset for the same set of operating points (i.e., BPCERs $\geq$ 32.25% for any APCER $\leq$ 10%). It is worth noting that a significant number of AP attempts were previously rejected by the biometric system, thereby resulting in the curve discontinuities in Fig. 5-b.



(a) RM-CSMAD. (b) CSMAD-RM.

Fig. 6: Score distributions computed by the PAD algorithm for the cross-database protocol.

3.3 Impact on Generalisation Capability

Finally, we evaluate the impact of the BR in the generalisation capability of the PAD approach over a cross-database protocol: one database is used for training of the AP detector while the remaining dataset is employed for testing. To simulate a real application, the thresholds of the BR are only tailored using the dev set in the test database evaluated. Fig. 6 reports the score distributions computed by the PAD approach for two train-test configurations. As it may be noted, score distributions for PAIs completely overlap with the one depicted by BPs. Despite a high number of AP attempts being previously rejected by the BR for CSMAD, we note in Fig. 6-a that the generalisation capability directly depends on the PAD method and not on the BR at hand, thus resulting in similar detection performances: BPCERs $\geq 95\%$ for APCERs $\leq 10\%$ are attained over different sets. Similar results can be observed for the CSMAD-RM configuration.

4 Conclusions

In this work, we analyse the impact of a BR on the overall PAD performance of a biometric system. In particular, we evaluate the facial BR module ArcFace together with a top state-of-the-art PAD method. The experimental evaluation conducted on well-known databases reported the feasibility of using first the BR to filter those samples considered both as a pristine. This configuration results in a detection performance improvement of the PAD approach: only 1% APs are accepted while at most 5% of BPs are rejected by the PAD scheme. Finally, we noted that the PAD generalisation capability in a cross-database evaluation directly depends on the PAD mechanism and not on the BR at hand. In order to extend our work, we plan in future directions to evaluate both proposed schemes over different biometric characteristics, biometric recognition systems, and PAD modules.

Acknowledgements

This research work has been funded by the DFG-ANR RESPECT Project (406880674) and the German Federal Ministry of Education and Research and the Hessen State Ministry for Higher Education, Research and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

References

[CAM13] Chingovska, I.; Anjos, A.; Marcel, S.: Anti-spoofing in action: joint operation with a verification system. In: Proc. Intl. Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 98–104, 2013.

[Ch19] Chingovska, I.; Mohammadi, A.; Anjos, A.; Marcel, S.: Evaluation methodologies for biometric presentation attack detection. In: Handbook of biometric anti-spoofing, pp. 457–480. 2019.

[Co16] Costa-Pazo, A.; Bhattacharjee, S.; Vazquez-Fernandez, E.; Marcel, S.: The REPLAY-MOBILE Face Presentation-Attack Database. In: Proc. Intl. Conf. on Biometrics Special Interests Group (BIOSIG). 2016.

[De19] Deng, J.; Guo, J.; Xue, N.; Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proc. Intl. Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 4690–4699, 2019.

[GSGBB21] Gonzalez-Soler, L. J.; Gomez-Barrero, M.; Busch, C.: On the Generalisation capabilities of Fisher Vector based Face Presentation Attack Detection. IET Biometrics, 10(5):480–496, September 2021.

[IAM14] I, Chingovska; Anjos, A.; Marcel, S.: Biometrics evaluation under spoofing attacks. IEEE Trans. on Information Forensics and Security (TIFS), 9(12):2264–2276, 2014.

[IS17] ISO/IEC JTC1 SC37 Biometrics: . ISO/IEC 30107-3. Information Technology - Biometric presentation attack detection - Part 3: Testing and Reporting. International Organization for Standardization, 2017.

[IS21] ISO/IEC JTC1 SC37 Biometrics: . ISO/IEC 19795-1:2021. Information Technology – Biometric Performance Testing and Reporting – Part 1: Principles and Framework. International Organization for Standardization, June 2021.

[MDR12] Marasco, E.; Ding, Y.; Ross, A.: Combining match scores with liveness values in a fingerprint verification system. In: Proc. Intl. Conf. on Biometrics: Theory, Applications and Systems (BTAS). pp. 418–425, 2012.

[MR14] Marasco, E.; Ross, A.: A survey on antispoofing schemes for fingerprint recognition systems. ACM Computing Surveys (CSUR), 47(2):1–36, 2014.

[Ra19] Ramachandra, R.; Venkatesh, S.; Raja, K.; Bhattacharjee, S.; Wasnik, P.; Marcel, S.; Busch, C.: Custom silicone face masks: Vulnerability of commercial face recognition systems & presentation attack detection. In: Proc. Intl. Workshop on Biometrics and Forensics (IWBF). pp. 1–6, 2019.