VULNERABILITY OF FACE AGE VERIFICATION TO REPLAY ATTACKS

Pavel Korshunov[1], *Anjith George*[1], *Gökhan Özbulak*[1], *and Sébastien Marcel*[1,2]

[1]Idiap Research Institute, Martigny, Switzerland
[2]University of Lausanne, Switzerland
{pavel.korshunov,anjith.george,gokhan.ozbulak,sebastien.marcel}@idiap.ch

ABSTRACT

Presentation attacks on biometric systems have long created significant security risks. The increase in the adoption of age verification systems, which ensure that only age-appropriate content is consumed online, raises the question of vulnerability of such systems to replay presentation attacks. In this paper, we analyze the vulnerability of face age verification to simple replay attacks and assess whether presentation attack detection (PAD) systems created for biometrics can be effective at detecting similar attacks on age verification. We used three types of attacks captured with iPhone 12, Galaxy S9, and Huawei Mate 30 phones from iPad Pro, which replayed the images from a commonly used UTKFace dataset of faces with true age labels. We evaluated four state of the art face age verification algorithms, including simple classification, distribution-based, regression via classification, and adaptive distribution approaches. We show that these algorithms are vulnerable to the attacks, since the accuracy of age verification on replayed images is only a couple of percentage points different compared to when the original images are used, which means an age verification system cannot distinguish attacks from bona fide images. Using two state of the art presentation attack detection systems, DeepPixBiS and CDCN, trained to detect similar attacks on biometrics, we demonstrate that they struggle to detect both: the types of attacks that are possible in age verification scenario and the type of bona fide images that are commonly used. These results highlight the need for the development of age verification specific attack detection systems for age verification to become practical.

Index Terms— Age verification, replay attacks, vulnerability, presentation attack detection, age anti-spoofing

1. INTRODUCTION

It is well known that presentation attacks (PAs) pose serious security risks to biometric systems [2, 3, 4]. Even such easy-to-perform attacks, like replaying a video or displaying a photo to the camera, can easily spoof state of the art face recognition [5]. In the past few years, a lot of research effort was therefore put into the development of presentation attack detection (PAD) systems and their integration with biometrics [6, 7, 8, 9, 10].

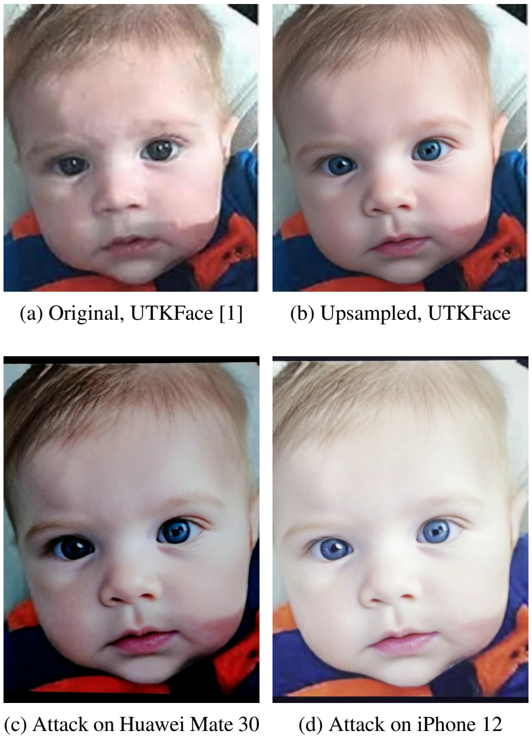

(a) Original, UTKFace [1]
(b) Upsampled, UTKFace
(c) Attack on Huawei Mate 30
(d) Attack on iPhone 12

Fig. 1: Examples of bona fide faces and attacks.

Age assurance or age verification systems are starting to become more and more prevalent, because they offer a solution to some of the oldest and widespread problems of misuse or disservice of an online content, like showing adult ads to minors, children having an unrestricted access to pornography, and adult predators pretending to be young kids in children video games. Therefore, recent years saw key legislation initiatives like the age appropriate design code[1] in the

This work is mainly funded by Swiss center for biometrics research and testing with partial contribution from InnoSuisse 108.708 INT-ICT project.

[1] https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/childrens-information/childrens-code-guidance-and-resources

UK , California Age Appropriate Design Code[2] and 144 state bills[3] in the US that will require internet companies to verify the age of people consuming their content.

However, for age verification systems to be practical, it is important to ensure their basic security and the ability to withstand at least such simple presentation attacks like replay attacks. However, the vulnerability of age verification systems to these attacks is not currently understood, at least, to our knowledge, no scientific study exists on the matter.

A possible way to defence against presentation attacks on age verification would be to adopt already existing (PAD) systems developed for biometrics. However, despite the obvious similarities, there are important differences that make the attacks on age verification harder to detect. For instance, for age verification, it is not necessary to preserve an identity of a subject, only a different age of a person need to be emulated. It means de-aging/aging [11] or other AI-based filters (depending on the purpose of a spoofing attempt), common in social media apps (TikTok, Snapchat, Instagram, etc.), can be used to change the appearance of a face to make it look younger or older. An attack may even present a photo of a younger or an older person and the photo does not need to represent anyone in particular. It means it is easier to perform attacks on age verification, since one can use any photos from internet to perform an attack. Also, age verification systems are built to detect mostly children [12], while children data is practically absent in the datasets on which PAD systems designed for biometrics are trained on.

In this paper, we consider the problem of presentation replay attacks on a face age verification system. To showcase the issue, we first built a dataset, referred to as UTKPAD[4], of the replay attacks by using the images from the *de facto* benchmark dataset UTKFace [1] used in face age verification evaluations. We took the photos from UTKFace, upsampled them with CodeFormer [13], and presented them on iPad Pro to three different mobile phones: iPhone 12, Galaxy S9, and Huawei Mate 30 (see examples in Figure 1). In this assumed scenario, an attacker uses an iPad to spoof an age verification system running on a mobile device, such as the one developed by Privately[5]. We have added the upsampling step, to an otherwise typical process of capturing replay attacks, to demonstrate the important difference between age verification and face recognition systems: various AI-based filters and upsampling methods, which do not guarantee the preservation of identities, can be used to attack age verification but may not work on face recognition.

By using the recaptured photos of the original UTKFace dataset, we assess several age verification approaches (see the details and comparison of these algorithms in [14]), including a baseline MobileNetV2-based classification approach and state of the art approaches that are based on regression via classification (RVC) training strategy [15], distribution training strategy [16], and recently proposed by us adaptive distribution strategy [14] on how vulnerable they are to the replay attacks.

To understand the feasibility of using the existing PAD systems already developed for biometrics, we evaluated two state of the art DeepPixBiS [17] and CDCN [18] approaches, trained on the OULU-NPU database [19] of presentation attacks. We tested these PAD approaches on both original and replayed images form UTKFace dataset to understand how well these systems can detect fake recaptured images from the original ones. Therefore, this paper aims to highlight the importance of the problem of presentation attacks on other systems besides biometrics, such as age verification, and to demonstrate if the existing detection solutions can be easily adapted to a new domain.

To allow researchers to verify, reproduce, and extend our work, we provide the attack protocols, evaluation scores, and the code to run and evaluate them as an open-source Python package[6].

2. UTKPAD: DATABASE FOR REPLAY ATTACKS

To create replay attacks of UTKPAD dataset[4] , we used 3504 images from the $eval$ set of UTKFace dataset [1], which contains face images with precise true age labels from a few months old babies all the way to 116 years old. All images in the database are labeled by age, gender, and ethnicity. The images cover large variation in pose, facial expression, illumination, occlusion, and resolution.

Before re-capturing the images, we have digitally upsampled them using pretrained CodeFormer [13] model, which was developed for enhancing and restoration of facial images. Since many of the original UTKFace images are of low resolution and were scraped from Internet, such upsampling improves visual quality of the attack. And since age verification does not require preservation of identities, any digital filter that outputs a realistic looking face of a desired age will pass as valid by a typical age verification system.

After upsampling, each image is converted into a 4-second video clip and all concatenated clips are then played on iPad Pro to iPhone 12, Galaxy S9, and Huawei Mate 30 (see examples in Figure 1) mobile phones. The final replay attack image is obtained by taking the middle frame from each corresponding segment of the recaptured video.

3. VULNERABILITY OF AGE VERIFICATION

It is important to note that an age verification can be considered as either a classification or regression problem, which

[2] https://californiaaadc.com/

[3] https://techpolicy.press/144-state-bills-aim-to-secure-child-online-safety-as-congress-flounders

[4] Database: https://www.idiap.ch/en/scientific-research/data/utkpad

[5] https://www.privately.eu/

[6] Source code: https://gitlab.idiap.ch/biometric/icassp2024.agepad

dictates the way the model is trained and evaluated. If the goal is to detect an actual age of a person, then regression is a logical choice, although age categories are more often used in practical applications, when the focus is on detecting children vs adults. In this paper, we define seven age categories: the person is a child (below 8 years old), of a puberty age (between 8 and 13), an adolescent (between 13 and 18), a young adult (between 18 and 25), an adult (between 25 and 35), of a middle age (between 35 and 50), and a senior (above 50) [20].

To evaluate the accuracy of the age verification as a categorical problem, we use *f1-score*, which is defined as f1-score $= \frac{2(P*R)}{P+R}$, where P precision and R is recall. The *f1-score* allows us to compare two different classifiers in a balanced way. To ensure the balanced *f1-score* value for the unbalanced data (the number of samples in different age categories vary a lot), we used a weighted variant of the metric. Also note that the higher the *f1-score* value is the better.

As the main underlying architecture for face age verification, we use MobileNetV2 [21]. We choose this model for its practicality in mobile applications and because the type of architecture does not affect the vulnerability of the overall system (we demonstrate it also by using ResNet50 backbone architecture for a comparative sanity check evaluation). Moreover, in vulnerability analysis, we evaluate the relative difference in the system's performance when original/bona fide images are used vs the replayed images. We assess the vulnerability of four different age verification systems (see more details in [14]) that are based on the following training strategies:

- *classification*: A baseline classifier with a cross entropy loss. A fully-connected layer of size equal to the number of classes (seven) is added at the top of MobileNetV2 or ResNet50 architectures.

- *rvc*: Regression via classification training strategy (RVC) [15] when an age range is split into several sets of classes using sliding window. The network has several heads (fully-connected layers), one for each split. At the inference, the average of the expected values on the outputs from several network heads is taken and is considered to be the predicted age. During the evaluation, we check in which of the seven categories, the predicted age falls.

- *distribution*: Distribution based training strategy [16], where instead of using one-hot encoding for true labels, as it is in a typical classification, a normal distribution with a specified sigma is used. It means that the ground truth label instead of a strict class becomes a distribution with the center at that true label.

- *adaptive*: A variant of the distribution-based strategy proposed in [14], where the sigma of the normal distribution is not fixed but is dynamic, depending on the true age label.

Table 1: Age verification methods [14] evaluated (*f1-score*) on bona-fide (BF) images of UTKFace dataset and the replay attacks performed with iPhone 12, Galaxy S9, and Huawei.

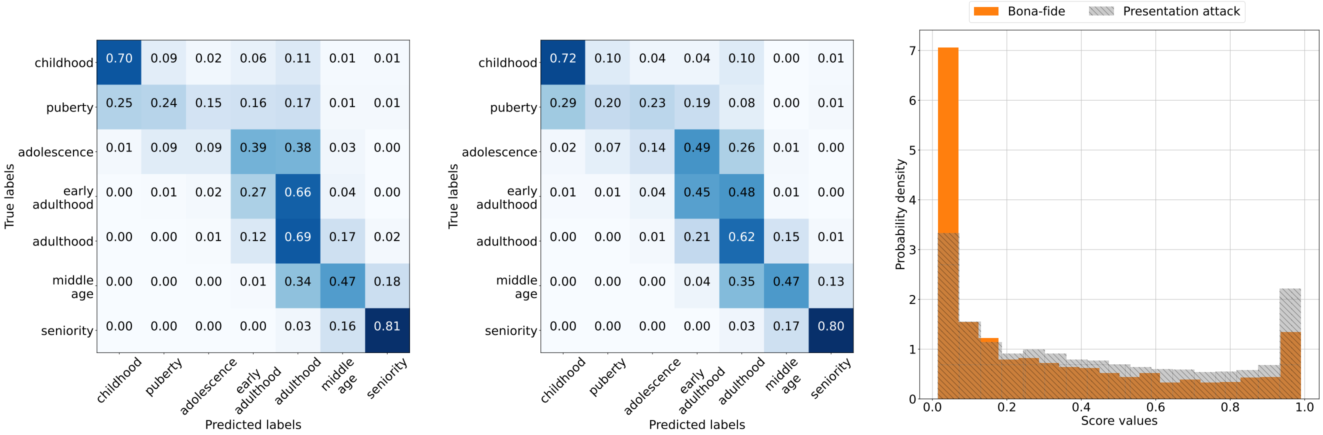
DB	Method	BF	iphone	galaxy	huawei
UTK	*adaptive*	0.599	0.566	0.567	0.586
ALL	*rvc*	0.596	0.571	0.573	0.583
ALL	*adaptive*	0.591	0.587	0.584	0.595
UTK	*class.*, ResNet50	0.591	0.540	0.561	0.561
ALL	*distribution*	0.589	0.574	0.585	0.597
UTK	*rvc*	0.581	0.534	0.554	0.573
UTK	*classification*	0.574	0.529	0.543	0.560
ALL	*classification*	0.567	0.516	0.510	0.536

4. PRESENTATION ATTACK DETECTION SYSTEMS

George and Marcel [17] introduced a system, DeepPixBiS, to use dense fully connected neural network architecture for presentation attack detection. This architecture was trained using pixel-wise binary supervision. By employing this form of supervision on the output maps, the neural network is compelled to develop shared representations that harness information from distinct patches across the face image. This reduces over-fitting and improved the cross dataset performance. DeepPixBiS is a commonly used baseline PAD system that demonstrates a state of the art performance in RGB-based presentation attack datasets.

Yu *et al.* [18] presents a PAD architecture leveraging central difference convolution (CDC). This architecture is designed for identifying intrinsic detailed patterns by aggregating both intensity and gradient information. Using this architecture, the authors proposed a CDC network model (CDCN) and its extension, CDCN++, which also employs and additional neural architecture search over a CDC search space and is integrated with the multiscale attention fusion module. Since CDCN++ shows better accuracy of attack detection, we used this version in our PAD evaluations.

Both of the PAD systems, DeepPixBiS and CDCN++, were trained using Protocol 1 of the OULU-NPU dataset. We evaluate PAD systems using metrics defined in the standard [22], Attack Presentation Classification Error Rate (APCER), Bonafide Presentation Classification Error Rate (BPCER), and Average Classification Error Rate (ACER) on $eval$ set of UTKFace with attacks of UTKPAD[4]. We compute APCER value using three different thresholds set on dev set of OULU dataset: i) when APCER is at 20% (denoted as BPCER5), ii) when APCER is at 5% (BPCER20), and iii) when the BPCER is equal to APCER, which is called equal error rate (EER). For the results on OULU dataset, we report ACER using the maximum $APCER$ of the four types of attacks (2 types of print and 2 type videos) ($APCER_{AP}$) following the ISO standard as $ACER = (max(APCER_{AP}) + BPCER)/2$.



(a) *adaptive* model on original UTKFace

(b) *adaptive* model on attacks by Huawei phone

(c) DeepPixBiS PAD performance

Fig. 2: Confusion matrices for *adaptive* age verification tested on bona fide UTKFace and on attacks by Huawei phone, plus, the score distribution of DeepPixBiS PAD in bona fide vs Huawei attacks scenario.

Table 2: The model performance in OULU-NPU Protocol 1.

	ACER ↓ (EER)	ACER ↓ (BPCER5)	ACER ↓ (BPCER20)
DeepPixBiS [17]	2.1	20.0	9.6
CDCN++ [18]	7.5	8.3	6.2

Table 3: CDCN++ and DeepPixBiS PAD evaluated on UTK-PAD with threshold set on OULU-NPU *dev* set.

Model	Replay attacks	ACER ↓ (EER)	ACER ↓ (BPCER5)	ACER ↓ (BPCER20)
DeepPixBiS [17]	iphone	38.9	40.1	37.3
	galaxy	48.7	52.4	50.2
	huawei	57.7	59.9	59.4
CDCN++ [18]	iphone	45.4	34.8	42.9
	galaxy	52.9	51.9	53.3
	huawei	61.2	61.3	63.3

5. EVALUATION RESULTS

We assess vulnerability of four age verification systems by comparing their performance, *f1-scores*, on the bona fide images of UTKFace dataset and the three replay attack of UTK-PAD[4] captured by iPhone 12, Galaxy S9, and Huawei Mate 30. We also evaluate whether two pretrained presentation attack detection systems can detect the attacks we collected.

5.1. Vulnerability assessment of age verification

Table 1 compares the performance of four age verification models for bona fide (BF) images and three replay attacks. The methods were either pretrained on training set of UTK-Face dataset, denoted as 'UTK', or a combination of several datasets, denoted as 'ALL', (see details in [14]), as indicated in 'DB' column. The results demonstrate that the differences in *f1-scores* between a bona fide or any of the attack scenarios are insignificant, in some cases, the age category detection accuracy is even higher for attacks. For instance, *f1-scores* of 0.597 for attacks using Huawei Mate 30 phone and for *adaptive* method pretrained on 'ALL' datasets is higher than 0.589 for bona fide. This result is illustrated by Figure 2, where two confusion matrices, showing the results for each age category, demonstrate that the same *adaptive* system performs better on Huawei attacks for 'childhood' and 'adolescence' categories.

5.2. Evaluation of PAD systems

Table 2 show the performance of PAD systems on the OULU dataset, with reasonably low error rates, depending on the set thresholds. When we use the same systems, the same thresholds, and the same metrics, but change the evaluation set to our UTKPAD, the error rates become above 34% across the board for both systems, as illustrated by Table 3, often reaching value of a random choice of 50%. Histogram score distribution in Figure 2 demonstrates this trend clearly, showing that the scores for bona fide and attacks overlap completely.

6. CONCLUSION

In this paper, we have shown that state of the art face age verification systems are vulnerable to replay attacks and the existing presentation attack detection systems, which perform well for the attacks on face recognition, cannot distinguish the attacks on age verification. It is evident that the problem of attacks is critical for age verification and the approaches for detecting these specific attacks need to be developed, since the legislation is requiring wide employment of age verification systems for protection of the children online.

7. REFERENCES

[1] Z. Zhang, Y. Song, and H. Qi, “Age progression/regression by conditional adversarial autoencoder,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4352–4360.

[2] S. Marcel, M. Nixon, and S. Li, “Handbook of biometric anti-spoofing-trusted biometrics under spoofing attacks,” *Advances in Computer Vision and Pattern Recognition. Springer*, 2014.

[3] I. Chingovska, N. Erdogmus, A. Anjos, and S. Marcel, “Face recognition systems under spoofing attacks,” in *Face Recognition Across the Imaging Spectrum*. Springer, 2016, pp. 165–194.

[4] R. Ramachandra and C. Busch, “Presentation attack detection methods for face recognition systems: A comprehensive survey,” *ACM Computing Surveys (CSUR)*, vol. 50, no. 1, pp. 1–37, 2017.

[5] A. Costa-Pazo, S. Bhattacharjee, E. Vazquez-Fernandez, and S. Marcel, “The replay-mobile face presentation-attack database,” in *International Conference of the Biometrics Special Interest Group (BIOSIG)*, Nov. 2016.

[6] S. Purnapatra, N. Smalt, K. Bahmani, P. Das, D. Yambay, A. Mohammadi, A. George, T. Bourlai, S. Marcel, S. Schuckers *et al.*, “Face liveness detection competition (LivDet-Face) – 2021,” in *IEEE International Joint Conference on Biometrics (IJCB)*, Aug. 2021.

[7] S. Marcel, J. Fierrez, and N. Evans, *Handbook of Biometric Anti-Spoofing: Presentation Attack Detection and Vulnerability Assessment*. Springer Nature, 2023.

[8] P. Korshunov and S. Marcel, “Impact of score fusion on voice biometrics and presentation attack detection in cross-database evaluations,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 695–705, June 2017.

[9] Y. Liu, A. Jourabloo, and X. Liu, “Learning deep models for face anti-spoofing: Binary or auxiliary supervision,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 389–398.

[10] Y. Atoum, Y. Liu, A. Jourabloo, and X. Liu, “Face anti-spoofing using patch and depth-based cnns,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 319–328.

[11] H. Pranoto, Y. Heryadi, H. L. H. S. Warnars, and W. Budiharto, “Recent generative adversarial approach in face aging and dataset review,” *IEEE Access*, vol. 10, pp. 28 693–28 716, 2022.

[12] EU Parliament, “Online age verification methods for children,” EU, 2023. [Online]. Available: https://www.europarl.europa.eu/RegData/etudes/ATAG/2023/739350/EPRS_ATA(2023)739350_EN.pdf

[13] S. Zhou, K. C. Chan, C. Li, and C. C. Loy, “Towards robust blind face restoration with codebook lookup transformer,” in *Advances in Neural Information Processing System (NeurIPS)*, 2022.

[14] P. Korshunov and S. Marcel, “Face anthropometry aware audio-visual age verification,” in *ACM International Conference on Multimedia (MM)*, Oct. 2022, pp. 5944–5951.

[15] A. Berg, M. Oskarsson, and M. O’Connor, “Deep ordinal regression with label diversity,” in *International Conference on Pattern Recognition (ICPR)*. IEEE, 2020, pp. 2740–2747.

[16] B.-B. Gao, H.-Y. Zhou, J. Wu, and X. Geng, “Age estimation using expectation of label distribution learning,” in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2018, pp. 712–718.

[17] A. George and S. Marcel, “Deep pixel-wise binary supervision for face presentation attack detection,” in *International Conference on Biometrics (ICB)*, 2019, pp. 1–8.

[18] Z. Yu, C. Zhao, Z. Wang, Y. Qin, Z. Su, X. Li, F. Zhou, and G. Zhao, “Searching central difference convolutional networks for face anti-spoofing,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5295–5305.

[19] Z. Boulkenafet, J. Komulainen, L. Li, X. Feng, and A. Hadid, “Oulu-npu: A mobile face presentation attack database with real-world variations,” in *IEEE international conference on automatic face & gesture recognition (FG)*, 2017, pp. 612–618.

[20] A. Dyussenbayev, “Age periods of human life,” *Advances in Social Sciences Research Journal*, vol. 4, no. 6, Apr. 2017.

[21] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, “MobileNets: Efficient convolutional neural networks for mobile vision applications,” *ArXiv*, vol. abs/1704.04861, 2017.

[22] ISO/IEC JTC 1/SC 37 Biometrics, “Information technology — International Organization for Standardization,” International Organization for Standardization, ISO Standard, Feb. 2016.