On the Detection of GAN-based Face Morphs using Established Morph Detectors

Luca Debiasi[1], Naser Damer[2,3], Alexandra Moseguí Saladié[3], Christian Rathgeb[4], Ulrich Scherhag[4], Christoph Busch[4], Florian Kirchbuchner[2], and Andreas Uhl[1]

[1] University of Salzburg, Salzburg, Austria
{ldebiasi,uhl}@cs.sbg.ac.at
[2] Fraunhofer Institute for Computer Graphics Research IGD, Darmstadt, Germany
{naser.damer,florian.kirchbuchner}@igd.fraunhofer.de
[3] TU Darmstadt, Darmstadt, Germany
alexamosegui93@gmail.com
[4] Hochschule Darmstadt, Darmstadt, Germany
{ulrich.scherhag,christian.rathgeb,christoph.busch}@h-da.de

Abstract. Face recognition systems (FRS) have been found to be highly vulnerable to face morphing attacks. Due to this severe security risk, morph detection systems do not only need to be robust against classical landmark-based face morphing approach (LMA), but also future attacks such as neural network based morph generation techniques. The focus of this paper lies on an experimental evaluation of the morph detection capabilities of various state-of-the-art morph detectors with respect to a recently presented novel face morphing approach, MorGAN, which is based on Generative Adversarial Networks (GANs).
In this work, existing detection algorithms are confronted with different attack scenarios: known and unknown attacks comprising different morph types (LMA and MorGAN). The detectors' performance results are highly dependent on the features used by the detection algorithms. In addition, the image quality of the morphed face images produced with the MorGAN approach is assessed using well-established no-reference image quality metrics and compared to LMA morphs. The results indicate that the image quality of MorGAN morphs is more similar to bona fide images compared to classical LMA morphs.

Keywords: Face Morphing · Generative Adversial Networks · Presentation Attack Detection

1 Introduction

Recently, automated face recognition systems (FRSs) are increasingly being used in different application scenarios, such as mobile device authentication or Automated Border Control (ABC). This wide spread deployment makes them attractive for attacks. In particular, their expected robustness to different environmental and user-specific conditions, e.g. varying illumination and subject poses,

and the widespread use of deep neural networks in FRS has been found to increase their vulnerability against presentation attacks [14]. In this context, face morphing attacks have attracted notable interest from the research community in the recent past.

Ferrara *et al.* [6] unleashed the vulnerability of FRSs against attacks based on morphed face images, which can be introduced in the issuance process of electronic travel documents due to security gaps. They compared morphed images with images of the original subjects using two commercial face recognition solutions, and concluded with the high vulnerability of face recognition to such attacks. Further studies considered the human expert vulnerability to morphed face images when comparing faces [7, 20]. They found out that human experts fails most of the times in detecting morphing attacks.

Different solutions were developed to detect face morphing attacks. Ramachandra *et al.* [19] were first to propose the automated detection of morphed face images. They applied local image descriptors such as the Binarised Statistical Image Features (BSIF) that capture textural properties of the image, which are later classified using a Support Vector Machine (SVM). Later works looked into using convolutional neural network(CNN) based features [18], image quality measures [16], the effect of printing and re-scanning the images [23], and differences between triangulating and averaging the facial landmarks on the detection [17]. Recent works by Debiasi *et al.* [4] propose to exploit the Photo Response Non-Uniformity (PRNU) of an image sensor to detect morphed face images, which is a widely used tool in the field of Digital Image Forensics (e.g. image forgery detection).

A standardised manner to evaluate the vulnerability of biometric systems to morphing attacks was recently proposed by Scherhag *et al.* [22]. A recent work by Ferrara *et al.* [8] viewed the morphing attack detction problem from a different perspective by proposing an approach to revert the morphed face image (demorph) enough to reveal the identity of the legitimate document owner, given a bona fide capture.

Other works considered that it might be possible in practice to use a live probe image along with the investigated image to detect a morphing attacks. This was done either by looking at the differential vector between both images [24], analysing the absolute distances and angles of the landmarks in both images [21], analysing the directed distances between these landmarks [1], or using the live probe image for demorphing [8]. The mentioned works so far developed and evaluated their approaches based on morphing attacks databases that were created based on facial landmarks.

Recently, a work by Damer *et al.* [2] proposed a new possibility of morphing attacks. They built their solution on generative adversarial networks (MorGAN). They morphed the latent representation of the morphed images and generated the morphing attacks based on that morphed latent vector. These morphing attacks proved to be hard to detect in the cases where they were not considered in the training process of the morphing detector [2].

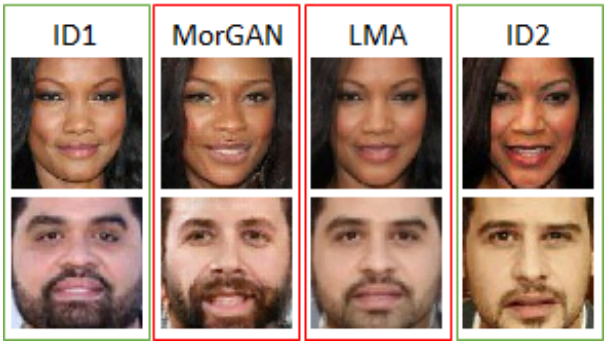


Fig. 1. Examples of the used morphing attacks, both the MorGAN and LMA. Original reference images are on the right and left.

The work presented in this paper aims at evaluating the detectability of LMA- and GAN-based morphed face images in different attack scenarios (known and unknown attacks) using several state-of-the-art morph detectors based on different features. The experimental evaluation performed in this work gives a preliminary outlook on the detectability future face morphing attacks. These attacks might include novel morphing strategies such as GANs for face morph generation, where it is not clear how the morph detection performance is affected by the artefacts that they introduce. For example, it is not clear if the properties of the image's PRNU are preserved in morphed images generated using a GAN-based approach or if the properties are altered, which has a decisive impact on the detection performance of PRNU-based morph detection approaches. Furthermore, this work also includes an image quality assessment of morphed face images generated using the MorGAN approach compared to classical LMA morphs.

The paper is organised as follows: the MorGAN approach and data set are described in Section 2. The image quality assessment of the generated MorGAN images is reported in Section 3, while the experimental setup and investigated state-of-the-art morph detectors are described in Section 4. The experimental results are reported and discussed in Section 5 and the paper is concluded in Section 6.

2 MorGAN Dataset

A database containing attacks created by the conventional landmark-based morphing technique, as well as the recently MorGAN-based approach, is used in this work. This allows the evaluation of detection performance of known and unknown attacks of the investigated morph detection approaches.

The database is based on recent work by Damer *et al.* [2] foreseeing using GANs to create morphing attacks and built on the CelebA [12] data set.

The MorGAN database contains a total of 1500 bona fide references, 1500 bona fide probes, 1000 LMA morphing attacks, and 1000 MorGAN morphing attacks. The database is split into disjoint (identity and image) and equal train and test sets, each including 750 bona fide references, 750 bona fide probes, and 500 attack images from each of both attack types (LMA and GAN). Because of computational and structural limitations of the MorGAN approach, the MorGAN attack images are of 64×64 pixels size (below the ICAO recommendations). Examples of the resulting image attacks and the original images creating these attacks are presented in Figure 1.

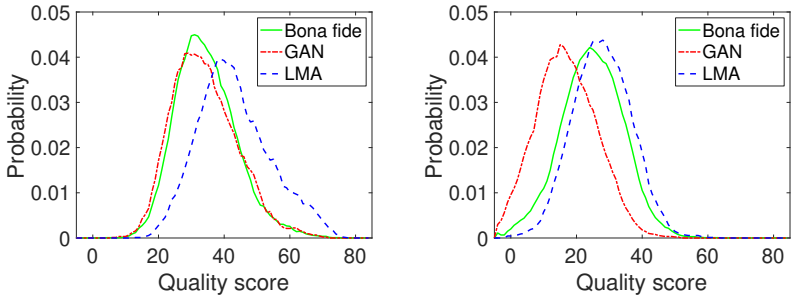
3 Quality of Morphed Face Images

As shown in [2] by Damer *et al.* , the morphed face images contained in the MorGAN data set are capable of successfully attacking pre-trained FRS, i.e. OpenFace and VGG-Face. They conclude that MorGAN attacks are weaker than the LMA ones, however, still make successful attacks on both FRSs. It has to be noted that the MorGAN approach has only recently been presented and that images with higher quality and resolution are expected to be generated with future versions of the approach.

In this work, the insights on the vulnerability of FRSs against face morph presentation attacks are complemented by an image quality analysis of the MorGAN morphs, which is compared to the quality of bona fide images and LMA morphs. Ferrara *et al.* [6] demonstrated, that even human experts are not able to discriminate between bona fide and high quality morphed face images. Therefore, the image quality of morphed plays an important role, since common pattern recognition techniques and humans in particular can easily detect obvious artefacts within the images. For examples on such obvious artefacts, the reader is referred to [22]. In order to assess the image quality of the different images in the MorGAN data set (bona fide, MorGAN and LMA morphs), the following no-reference image quality metrics have been evaluated on all 1500 bona fide, 1000 MorGAN and 1000 LMA images: BIQI [15], BRISQUE [13], OG-IQA [10] and SSEQ [11].

To render a fair comparison with the MorGAN images possible, LMA and bona fide images have been downsized to the same resolution of 64×64 pixels. We did not consider any face-specific sample quality assessment metrics in this work due to the small resolution of the MorGAN images.

All image quality results are illustrated in Table 1, while only two selected quality metrics are presented in Figure 2. Overall, the evaluation shows that the image quality of both morphed MorGAN and LMA images is very similar to the image quality of the bona fide images within the MorGAN data set. BIQI, OG-IQA and SSEQ show that the image quality score distributions of MorGAN images are more resemblant of the bona fide distribution compared to LMA morphs. Only BRISQUE shows a different result, where the quality scores of LMA morphs are more alike the ones of bona fide images compared to MorGAN morphs.



(a) BIQI (b) BRISQUE

Fig. 2. Image quality score distributions of bona fide images compared to LMA and MorGAN-based morphs.

These results, using equally sized images of 64×64 pixels, reveal that morphed images generated with the MorGAN approach are are more similar to bona fide images compared to the classical LMA approach in respect to their image quality, which is underlined by the distortion independence (BIQI), generalisability (OG-IQA) and closeness to human perception (SSEQ) of the image quality metrics supporting these results.

4 Experimental Setup

This study aims at investigating the detection performance of various morph detection approaches based on distinct features for MorGAN attacks. In particular, their ability of dealing with known and unknown attacks is of special interest, especially when future attacks based on unknown (neural network based) morphing techniques are considered.

4.1 Morph Detection Algorithms

Our morph attack detection methodology aims at enabling a wider range of conceptual evaluation and more diverse coverage of the state-of-the-art by considering image feature extraction methods of three different natures. One is the hand crafted classical image descriptors, the Local Binary Pattern Histogram (LBPH) [18], the second is based on transferable deep-CNN features [19] and the third type is based on the Photo Response Non-Uniformity (PRNU) [3, 4]. All three types of features were previously utilised for the detection of face morphing attacks based on LMA approaches.

4.2 Experiments

The morph attack detection experiments are ordered by the feature type (CNN, LBPH, PRNU-VAR and PRNU-HIST) and by the type of attack, i.e. known

Table 1. Statistical properties of image quality metrics for bona fide images and LMA and MorGAN-based morphed images.

Metric	Property	Bona fide	MorGAN	LMA
BIQI	Mean	35.06	34.56	43.55
	Std	8.95	9.51	10.71
	Min	8.43	10.83	17.47
	Max	71.86	67.13	73.17
BRISQUE	Mean	25.22	17.23	28.30
	Std	9.13	9.45	8.50
	Min	-3.31	-12.71	2.28
	Max	59.76	90.29	59.29
OG-IQA	Mean	-0.82	-0.87	-0.74
	Std	0.09	0.07	0.10
	Min	-0.95	-0.95	-0.94
	Max	-0.25	-0.39	-0.39
SSEQ	Mean	30.25	29.51	37.80
	Std	9.30	7.82	7.70
	Min	-6.78	4.71	3.48
	Max	59.76	55.81	62.26

or unknown and the type of morphs used for the attack (MorGAN and LMA). Due to the nature of the investigated detection algorithms and their design, the experiments had to be conducted in a slightly different manner for the various detectors, in order to ensure fair and comparable results. This has an effect on the sample size used for evaluation and the number of unknown attacks, which is described in more detail in the following.

Since CNN and LBPH are learning-based algorithms, the data is split into distinct train and test sets, both containing 750 bona fide images and 500 images for each attack type (LMA and MorGAN). A "known" attack (K) is given when the algorithm is evaluated with the same attack type as it is trained with, e.g. the algorithm was trained using LMA morphs and is evaluated on LMA morphs. An "unknown" attack (U), on the other hand, is given when different attack types are used to train and evaluate the algorithm, e.g. the algorithm is trained using LMA morphs and evaluated on MorGAN morphs. This leads to the following attack types for CNN and LBPH:

- K-LMA: Trained with LMA morphs, tested with LMA morphs.
- K-MorGAN: Trained with MorGAN morphs, tested with MorGAN morphs.
- U-LMA: Trained with MorGAN morphs, tested with LMA morphs.
- U-MorGAN: Trained with LMA morphs, tested with MorGAN morphs.

The two PRNU-based algorithms, PRNU-VAR and PRNU-HIST, do not rely on any training for classification, thus the whole data set, comprised of 1500 bona fide images and 1000 images for each attack type (LMA and MorGAN), is used for evaluation of the detectors. Therefore, all attacks with LMA or MorGAN

morphs can be considered as "unknown" (U) for the PRNU-based algorithms. This leads to the following attack types for PRNU-VAR and PRNU-HIST:

- U-LMA: Tested with LMA morphs.
- U-MorGAN: Tested with MorGAN morphs.

4.3 Evaluation

The assessment of the morph detection performance is based on metrics defined in ISO/IEC 30107-3 [9]: Attack Presentation Classification Error Rate (APCER) and Bona Fide Presentation Classification Error Rate (BPCER), as suggested in literature [22]. APCER defines the proportion of morphed face presentations incorrectly classified as bona fide presentations, while BPCER is the proportion of bona fide presentations incorrectly classified as morphed face presentation attacks. The detection systems are evaluated at different operating points: The operation point of the system, where APCER = BPCER, is defined as detection equal error rate D-EER. Furthermore, two additional operation points, BPCER10 (where APCER = 10%) and BPCER20 (where APCER = 5%), are reported.

5 Morph Detection Results

The outcome of the morph detection experiments, conducted according to Section 4, are summarised in Table 2 and illustrated with DET plots in Figure 3.

Table 2 shows the D-EER, BCPER10 and BCPER20 results for the various attack scenarios and morph detection algorithms described in Section 4. CNN shows the best performance at detecting LMA morphs, independent of the attacks being known or unknown. It achieves a perfect result for the K-LMA attack, and a D-EER of only 4% for U-LMA. However, it struggles in case of K-MorGAN or completely fails to detect U-MorGAN attacks. LBPH yields the overall lowest error rates among all morph detection algorithms and across all attack scenarios. It is able to detect both LMA and MorGAN morphs, but the performance gap between known and unknown attacks is very large. For known attacks, it is able to achieve low D-EERs of 9% for LMA and 1% for MorGAN attacks, while for unknown attacks the performance drops significantly to 23% and 19%, respectively. The results indicate that the CNN and LBPH detectors are not able to generalise well over different attack types, as it can be clearly seen in Figures 3(a) and 3(b), which might be caused by the closed-set training design of both algorithms.

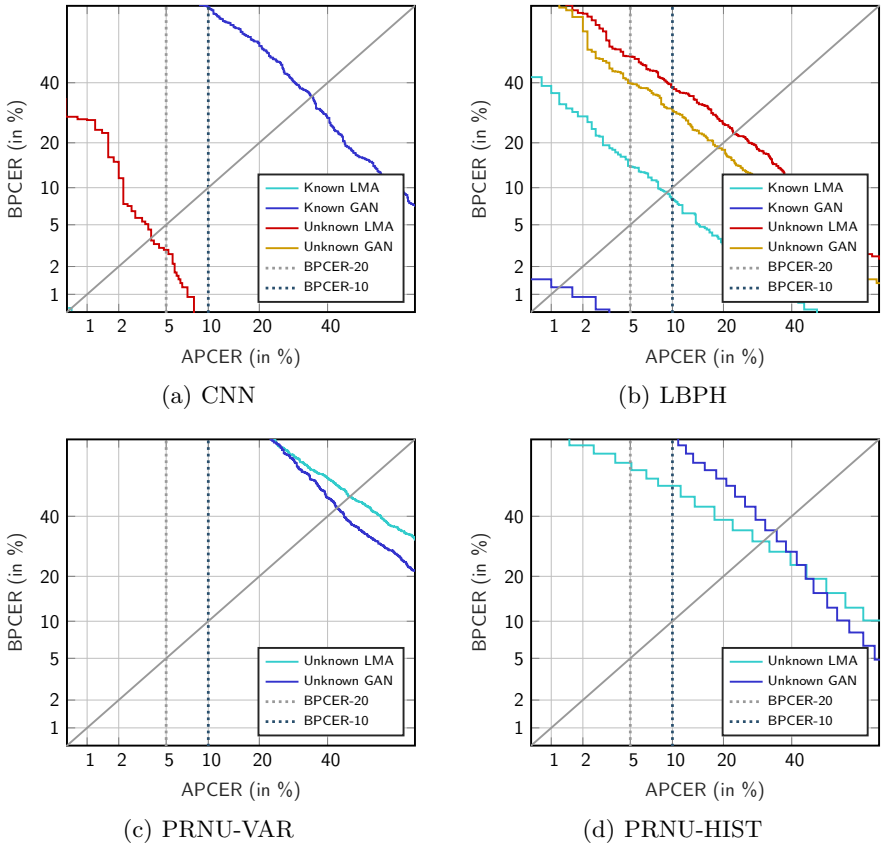
The performance of the two PRNU-based algorithms is worse compared to the previously discussed CNN and LBPH algorithms, with D-EERs around 45% for PRNU-VAR and 30% for PRNU-HIST. Nonetheless, the results for these two algorithms show a very promising property: their stable performance across all attack types (known and unknown) and morph types (MorGAN and LMA). This

Table 2. Morph detection performance of investigated algorithms under different attack scenarios.

Algorithm	Attack Type	D-EER	BCPER10	BCPER20
CNN	K-LMA	0.00	0.00	0.00
	K-MorGAN	0.34	0.67	0.78
	U-LMA	0.04	0.00	0.02
	U-MorGAN	0.50	0.90	0.95
LBPH	K-LMA	0.09	0.08	0.14
	K-MorGAN	0.01	0.00	0.00
	U-LMA	0.23	0.38	0.49
	U-MorGAN	0.19	0.29	0.39
PRNU-VAR	U-LMA	0.47	0.85	0.92
	U-MorGAN	0.43	0.85	0.92
PRNU-HIST	U-LMA	0.30	0.49	0.58
	U-MorGAN	0.33	0.69	0.81

consistency becomes evident when looking at Figures 3(c) and 3(d). While they might not perform as well as CNN and LBPH in some cases, the results indicate a high potential for the generalisabilty of PRNU-based algorithms across different morph types, independently of the morph type being known or unknown. Furthermore, it can be observed that the PRNU of MorGAN morphs shows similar properties as the PRNU of LMA-based morphs, which leads to an almost equal detection performance for the PRNU-based detectors. Due to time and space constraints, a more thorough investigation of the PRNU signal resulting from the GAN operations is left for future research, in particular whether a PRNU-based identification of the source camera in images generated with GANs might still be possible. The D-EER performance of the two approaches is reported to be much better for larger images (320x320 pixels) in [4] and [3], thus we conclude that the overall poor performance for the PRNU-VAR and PRNU-HIST is a result of the small image size of 64×64 pixels in the MorGAN data set. It is commonly known in the field of Digital Image forensics, that the performance of PRNU-based approaches tends to degrade significantly with smaller image resolutions, as it is shown in [5].

Summarising the morph detection results, it can be observed that all investigated detection algorithms have their advantages and drawbacks. CNN works well for detecting LMA attacks, but fails at detecting MorGAN attacks. LBPH works quite well overall, but shows a high performance gap between known and unknown attacks, leaving it vulnerable for unknown attacks. PRNU-HIST and PRNU-VAR show an overall weak performance (due to the small image resolution), but they have the big advantage of being very stable across all evaluated attacks. If the general performance of the PRNU-based algorithms can be improved, it can be expected that they will show a high robustness against many unknown attack scenarios.



(a) CNN (b) LBPH

(c) PRNU-VAR (d) PRNU-HIST

Fig. 3. DET plots for investigated morphing detection algorithms and different attack scenarios.

6 Conclusion

The detection of morphed face images has become an important part of automated face recognition systems, due to their severe vulnerability to such attacks.

In this work, we investigate the performance of different state-of-the-art face morph detection algorithms on the recently proposed MorGAN data set. This data set, besides containing bona fide images and classical landmark-based morphs, also contains morphed images generated using the MorGAN approach. As the name implies, this novel type of morphed face images is created using Generative Adversarial Networks. The focus of this work lies on the evaluation of different attack scenarios: known and unknown attacks as well as different morph types. Furthermore, we also compare the image quality of MorGAN images to LMA based morphs using different well-established no-reference image quality metrics to evaluate the quality of generated morphs. The experimental evaluation performed in this work gives a preliminary prospect at the detection

of future face morphing attacks, which might make use of unknown, most likely neural network based, morph generation techniques.

Summarising, the image quality assessment shows that the quality of MorGAN face morphs is closer to the quality of bona fide images as compared to classical LMA morphs, which underlines the capabilities of the MorGAN morph generation approach.

The morph detection performance results for the state-of-the-art detectors show that CNN fails at detecting the MorGAN morphs, but excels at detecting the classical LMA morphs. LBPH can achieve a very low D-EER of 1% for MorGAN and 9% for LMA morphs, but only in the case of known attacks. However, the performance of LBPH lacks consistency when confronted with unknown attacks. The two PRNU-based algorithms show a weaker overall performance of around 30% in the best case for both MorGAN and LMA morphs, which is most likely caused by the small image resolution.

Clearly, the MorGAN approach needs to be enhanced and further developed to produce images with higher resolutions, i.e. ICAO compliant images. This would allow for a more comprehensible analysis of the detectability and quality of the generated morphed face images.

References

1. Damer, N., Boller, V., Wainakh, Y., Boutros, F., Terhörst, P., Braun, A., Kuijper, A.: Detecting face morphing attacks by analyzing the directed distances of facial landmarks shifts. In: Proceedings of the 40th German Conf. on Pattern Recognition (GCPR 2018). Lecture Notes in Computer Science, Springer (2018)
2. Damer, N., Saladie, A.M., Braun, A., Kuijper, A.: MorGAN: Recognition vulnerability and attack detectability of face morphing attacks created by generative adversarial network. In: 9th IEEE Int. Conf. on Biometrics Theory, Applications and Systems (BTAS 2018). IEEE (2018)
3. Debiasi, L., Rathgeb, C., Scherhag, U., Uhl, A., Busch, C.: PRNU variance analysis for morphed face image detection. In: Proceedings of the IEEE 9th Int. Conf. on Biometrics: Theory, Applications, and Systems (BTAS 2018). pp. 1–8. Los Angeles, California, USA (10 2018)
4. Debiasi, L., Scherhag, U., Rathgeb, C., Uhl, A., Busch, C.: PRNU-based detection of morphed face images. In: 2018 Int. Workshop on Biometrics and Forensics (IWBF 2018). pp. 1–7 (6 2018)
5. Debiasi, L., Uhl, A.: Prnu enhancement effects on biometric source sensor attribution. IET Biometrics **6**(4), 256–265 (2017)
6. Ferrara, M., Franco, A., Maltoni, D.: The magic passport. In: IEEE Int. Joint Conf. on Biometrics (IJCB 2014). pp. 1–7. IEEE (2014)
7. Ferrara, M., Franco, A., Maltoni, D.: On the effects of image alterations on face recognition accuracy. In: Bourlai, T. (ed.) Face Recognition Across the Imaging Spectrum, pp. 195–222. Springer (2016)
8. Ferrara, M., Franco, A., Maltoni, D.: Face demorphing. IEEE Trans. Information Forensics and Security **13**(4), 1008–1017 (2018)
9. ISO/IEC JTC1 SC37 Biometrics: ISO/IEC IS 30107-3:2017, IT – Biometric presentation attack detection – Part 3: Testing and Reporting

10. Liu, L., Hua, Y., Zhao, Q., Huang, H., Bovik, A.C.: Blind image quality assessment by relative gradient statistics and adaboosting neural network. Signal Processing: Image Communication **40**, 1–15 (2016)
11. Liu, L., Liu, B., Huang, H., Bovik, A.C.: No-reference image quality assessment based on spatial and spectral entropies. Signal Processing: Image Communication **29**(8), 856–863 (2014)
12. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of Int. Conf. on Computer Vision (ICCV 2015) (12 2015)
13. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Transactions on Image Processing **21**(12), 4695–4708 (2012)
14. Mohammadi, A., Bhattacharjee, S., Marcel, S.: Deeply vulnerable: a study of the robustness of face recognition to presentation attacks. IET Biometrics **7**(1), 15–26 (2018)
15. Moorthy, A.K., Bovik, A.C.: A two-step framework for constructing blind image quality indices. IEEE Signal processing letters **17**(5), 513–516 (2010)
16. Neubert, T.: Face morphing detection: An approach based on image degradation analysis. In: Kraetzer, C., Shi, Y., Dittmann, J., Kim, H.J. (eds.) Proceedings of the 16th Int. Workshop on Digital Forensics and Watermarking (IWDW 2017). Lecture Notes in Computer Science, vol. 10431, pp. 93–106. Springer (2017)
17. Ramachandra, R., Raja, K.B., Venkatesh, S., Busch, C.: Face morphing versus face averaging: Vulnerability and detection. In: IEEE Int. Joint Conf. on Biometrics (IJCB 2017). pp. 555–563. IEEE (2017)
18. Ramachandra, R., Raja, K.B., Venkatesh, S., Busch, C.: Transferable deep-cnn features for detecting digital and print-scanned morphed face images. In: 2017 IEEE Conf. on Computer Vision and Pattern Recognition Workshops, CVPR. pp. 1822–1830. IEEE Computer Society (2017)
19. Ramachandra, R., Raja, K.B., Busch, C.: Detecting morphed face images. In: 8th IEEE Int. Conf. on Biometrics Theory, Applications and Systems (BTAS 2016). pp. 1–7. IEEE (2016)
20. Robertson, D.J., Kramer, R.S.S., Burton, A.M.: Fraudulent id using face morphs: Experiments on human and automatic recognition. PLOS ONE **12**(3), 1–12 (03 2017)
21. Scherhag, U., Budhrani, D., Gomez-Barrero, M., Busch, C.: Detecting morphed face images using facial landmarks. In: Mansouri, A., Elmoataz, A., Nouboud, F., Mammass, D. (eds.) Proceedings of the 8th Int. Conf. on Image and Signal Processing (ICISP 2018). Lecture Notes in Computer Science, vol. 10884, pp. 444–452. Springer (2018)
22. Scherhag, U., Nautsch, A., Rathgeb, C., Gomez-Barrero, M., Veldhuis, R., Spreeuwers, L., Schils, M., Maltoni, D., Grother, P., Marcel, S., Breithaupt, R., Raghavendra, R., Busch, C.: Biometric systems under morphing attacks: Assessment of morphing techniques and vulnerability reporting. In: Int. Conf. of the Biometrics Special Interest Group (BIOSIG 2017). pp. 1–12 (2017)
23. Scherhag, U., Ramachandra, R., Raja, K.B., Gomez-Barrero, M., Rathgeb, C., Busch, C.: On the vulnerability of face recognition systems towards morphed face attacks. In: 5th Int. Workshop on Biometrics and Forensics (IWBF 2017). pp. 1–6. IEEE (2017)
24. Scherhag, U., Rathgeb, C., Busch, C.: Towards detection of morphed face images in electronic travel documents. In: 13th IAPR Int. Workshop on Document Analysis Systems, (DAS 2018). pp. 187–192. IEEE Computer Society (2018)