Deep Face Representations for Differential Morphing Attack Detection

Ulrich Scherhag, Christian Rathgeb, Johannes Merkle, and Christoph Busch

Abstract—**The vulnerability of facial recognition systems to face morphing attacks is well known. Many different approaches for morphing attack detection (MAD) have been proposed in the scientific literature. However, the MAD algorithms proposed so far have mostly been trained and tested on datasets whose distributions of image characteristics are either very limited (e.g., only created with a single morphing tool) or rather unrealistic (e.g., no print-scan transformation). As a consequence, these methods easily overfit on certain image types and the results presented cannot be expected to apply to real-world scenarios. For example, the results of the latest NIST FRVT MORPH show that the majority of submitted MAD algorithms lacks robustness and performance when considering unseen and challenging datasets. In this work, subsets of the FERET and FRGCv2 face databases are used to create a realistic database for training and testing of MAD algorithms, containing a large number of ICAO-compliant bona fide facial images, corresponding unconstrained probe images, and morphed images created with four different face morphing tools. Furthermore, multiple post-processings are applied on the reference images, e.g., print-scan and JPEG2000 compression. On this database, previously proposed differential morphing algorithms are evaluated and compared. In addition, the application of deep face representations for differential MAD algorithms is investigated. It is shown that algorithms based on deep face representations can achieve very high detection performance (less than 3% D-EER) and robustness with respect to various post-processings. Finally, the limitations of the developed methods are analyzed.**

Index Terms—**Biometrics, face recognition, morphing attacks, morphing attack detection, differential attack detection, deep face representation.**

I. Introduction

IMAGE morphing techniques can be used to combine information from two (or more) images into one image. Morphing techniques can also be used to create a morphed facial image from the biometric face images of two individuals, of which the biometric information is similar to that of both individuals. Realistically looking morphed face images can be generated by unskilled users applying readily available tools [1]. An example of a morphed facial image is shown as part of Figure 1.

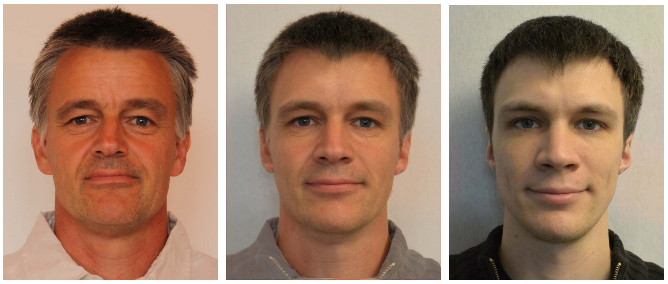

(a) Subject 1 (b) Morph (c) Subject 2

Fig. 1. Example for a morphed face image (b) of subject 1 (a) and subject 2 (c). The Morph was manually created using FantaMorph.

In many countries, the facial image submitted for an electronic travel document is provided by the applicant either in analogue (i.e., print on paper) or digital form. Therefore, an attacker (e.g., a wanted criminal or a foreigner without authorization to enter a territory) could morph his face image with the face image of a similar looking accomplice who could apply for a passport or another electronic travel document with the morphed image. Since many morphed images are similar enough to deceive human examiners as well as automatic face recognition systems [2], [3], the attacker can then use the electronic travel document issued to the accomplice to pass through automatic or manual border controls. The vulnerability of automated face recognition systems against such a Morphing Attack (MA) was initially showcased in [4]. The potential to launch a MA in practice was demonstrated by members of the political activist group *Peng! Kollektiv*, who succeeded without any problem in applying for a passport with a morphed face image.[1] Morphing attacks can be applied to other modalities as well. Whether a system is susceptible to this kind of attack can be theoretically examined with the framework presented in [5] and [6].

Many approaches for Morphing Attack Detection (MAD) algorithms have already been published in the scientific literature. Most of them can be categorized as *single image*

Manuscript received November 15, 2019; revised April 1, 2020 and April 27, 2020; accepted May 5, 2020. Date of publication May 14, 2020; date of current version July 6, 2020. This work was supported in part by the German Federal Ministry of Education and Research, in part by the Hessen State Ministry for Higher Education, Research and the Arts within their joint support of the ATHENE (National Research Center for Applied Cybersecurity), and in part by the Federal Office of Information Security (BSI) through the FACETRUST Project. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Clinton Fookes. *(Corresponding author: Ulrich Scherhag.)*

Ulrich Scherhag and Christoph Busch are with the da/sec—Biometrics and Internet Security Research Group, Hochschule Darmstadt, 64295 Darmstadt, Germany (e-mail: ulrich.scherhag@h-da.de).

Christian Rathgeb is with the da/sec—Biometrics and Internet Security Research Group, Hochschule Darmstadt, 64295 Darmstadt, Germany, and also with secunet Security Networks AG, 45138 Essen, Germany.

Johannes Merkle is with secunet Security Networks AG, 45138 Essen, Germany.

Digital Object Identifier 10.1109/TIFS.2020.2994750

[1] Peng! Kollektiv, MaskID: https://pen.gg/de/campaign/maskid/

MAD algorithms, which examine only the potentially morphed face image, as opposed to *differential MAD* methods that compare the image in question with a trusted probe image (e.g., a live capture). While many publications report impressive detection rates, these results are hardly applicable to real-world scenarios. Firstly, the datasets used for evaluation are not realistic. In particular, most publications ([18]–[20] and [21] being exceptions) do not consider variations in post-processings of the images, e.g., print-scan transformation or severe compression, which can occur in real-world scenarios and may drastically reduce detectable artefacts from the morphing process. In addition, the vast majority of previous publications on differential MAD (with the notable exception of [14]) use probe images that do not exhibit a realistic intra-subject variance, e.g., facial pose and expression, illumination, the subject's appearance (glasses, beard, hair style, clothing, cosmetics) and aging. Secondly, the datasets used for training and evaluation are not sufficiently distinct but contain images with similar characteristics. In particular, the test sets typically contain only morphed images generated with the same tools as the images in the training set. Consequently, the low error rates reported in these publications may simply reflect an overfitting to the specific and unrealistic properties of the images used. This suspicion is supported by the recent results of the Face Recognition Vendor Test (FRVT) MORPH test [22] conducted by the National Institute of Standards and Technology (NIST), where none of the submitted algorithms achieved satisfactory performance for all test scenarios.

The main contributions of this work are:

- A conceptual categorization of published MAD approaches including technical considerations and trade-offs, along with a more detailed survey of relevant differential MAD schemes.
- The creation of a face database containing a large number of ICAO-compliant bona fide and morphed reference images (created using four different morphing algorithms) as well as corresponding unconstrained probe images, which are obtained from different image sources, i.e., face image databases; the use of image post-processings which are likely to be applied in real world, i.e., JPEG2000 compression or print-scan transformations, together with the selection of probe images exhibiting variations in facial pose and expression as well as illumination and focus enable the training and testing of MAD algorithms in a realistic scenario.
- The proposal of differential MAD based on deep face representations; by employing state-of-the-art face recognition systems, rich and compact deep facial representations are extracted from pairs of reference and probe images and combined to train machine learning-based classifiers to detect image alterations induced by morphing algorithms.
- A comprehensive vulnerability assessment of state-of-the-art face recognition systems against MAs followed by a scenario-based evaluation of the proposed MAD concept applying a commercial and two open-source deep face recognition systems; it is shown that the MAD based on deep face representations outperforms previously proposed differential MAD schemes, which is also confirmed by the latest NIST FRVT MORPH report [22].

The remainder of this work is structured as follows: related works on MAD are summarized in Section II. In Section III we describe the databases set up for our investigations. Section IV describes our approach to develop MAD algorithms based on deep face representations. In Section V we evaluate our approach and compare the results with our evaluation of other state-of-the-art MAD algorithms. Finally, conclusions are drawn in Section VI.

II. Related Work

In recent years, numerous MAD approaches have been proposed. The following subsections give a rough overview of single image and differential MAD algorithms. A more detailed listing and description of the individual algorithms can be found in [1].

A. Single Image MAD

Single image MAD approaches can be categorized as *texture descriptors*, e.g., in [23], *forensic image analysis*, e.g., in [24], and methods based on *deep neural networks*, e.g., in [25]. These differ in the artefacts they can potentially detect.

Texture descriptors, e.g., Local Binary Patterns (LBP) [26] or Binarized Statistical Image Features (BSIF) [27], attempt to extract discriminative information from images, which can be employed for the purpose of texture classification. The morph process averages the images, which results in smoothed skin textures. Furthermore, ghost artefacts or half-shade effects can occur due to regions that do not overlap exactly (e.g., hair). In particular, in the area of the pupils and nostrils these artefacts appear more frequently, for examples the reader is referred to [28]. In addition, distorted edges or shifted image areas can occur. These types of artefacts can be easily represented and detected using texture descriptors in multiple ways [7], [20], [21], [23], [29]–[37].

Under the assumption that the morphing process leaves specific traces in the image, forensic image analysis techniques can be used to detect them. By averaging the images, the sensor pattern noise is changed. It was shown in [38]–[42] that these changes can be used for MAD. Under the assumption that the images are intermediately stored during the morph creation process employing lossy compression algorithms, double compression artefacts can be analyzed [17], [43]. Furthermore, inconsistencies in the image, e.g., inconsistent illumination [44] or color values, might be evaluated.

Deep neural networks can be used to detect morphs in two different ways. Firstly, a new neural network is trained from scratch or an existing neural network is re-trained [19], [25], [45] for the task of MAD. Deep neural nets can theoretically be trained to detect any artefact. Therefore, it is important that the training data contains a high variance, to avoid overfitting to algorithm and database specific artefacts. Secondly, the feature vectors (embeddings) extracted by existing deep nets can be employed for MAD [34]. Since the neural network was not

TABLE I

OVERVIEW OF MOST RELEVANT DIFFERENTIAL MAD ALGORITHMS

Reference	Approach	Category	Morphing method	Source face database	Post-processing	Remarks
Scherhag *et al.* [7]	Differences in BSIF features with SVM	Feature comparison	triangulation + blending	FRGCv2 [8]	–	–
Scherhag *et al.* [9]	Differences in angles of landmark pairs with SVM	Feature comparison	triangulation + blending	ARface [10]	–	–
Damer *et al.* [11]	Directed distances of landmarks with SVM	Feature comparison	triangulation + blending (+ swapping)	FERET [12]	–	–
Ferrara *et al.* [13]	Demorphing	Morphing reversion	triangulation + blending, GIMP/GAP	ARface [10]	–	–
Ferrara *et al.* [14]	Demorphing	Morphing reversion	triangulation + blending, GIMP/GAP	ARface [10], CAS-PEAL-R1 [15]	–	CAS-PEAL-R1 contains images with pose variations
Peng *et al.* [16]	GAN-based Demorphing	Morphing reversion	triangulation + blending [17]	in-house	–	–

trained on morphed facial images, it can be assumed that no overfitting to unrealistic morphing artefacts occurs.

B. Differential MAD

Published differential methods and their properties are listed in Table I. Differential MAD can be divided into two categories. The first category are algorithms that compare feature vectors extracted from extracted from trusted live captures and potential morphs. Single image algorithms can be extended to differential algorithms, e.g., by estimating differences between feature vectors [7]. The additional information of the trusted live capture might improve the performance and robustness of the detection algorithm. Further, algorithms explicitly utilizing this additional information have been introduced in [9], [11], where the distances between features of facial landmarks are estimated.

The second category contains algorithms that try to reverse the process of morphing. Here the assumption is that if two subjects are represented in the morphed image and the trusted live capture is subtracted from the possibly morphed reference, in the case of a morphed face image the identity of the second subject becomes more dominant, leading to a lower face recognition scores. If there is no other subject in the image, only the existing one remains. In [13], [14] the so-called demorphing algorithm was proposed, an approach where reversion is done explicitly. In addition to the explicit demorphing approach, demorphing based on a Generative Adversarial Network (GAN) is proposed in [16].

III. CREATION OF MAD DATABASE

For the development and evaluation of MAD algorithms, bona fide and morphed reference images are required. In order to resemble passport photos, these images should meet the requirements of the ICAO passport photo quality standards [46]. Multiple tools should be used to generate morphed images to represent a sufficient variance of MAs. In addition, the reference images (bona fide and morphed) should have undergone various realistic post-processings including strong JPG2000 compression and print-scan transformation. For the investigation of differential MAD, additional probe images are needed. In order to simulate the important use-case of automatic border control, theses probe images should resemble live-captures taken in eGates. Since these recordings are semi-controlled the quality of the captured samples is degraded and does not comply to the ICAO requirements. A much higher variance can be expected, e.g., with respect to pose, facial expression, illumination, and background. Furthermore, as the border control can occur up to ten years after the passport application, reference and probe images can significantly differ with respect to the subject's appearance (glasses, beard, hair style, clothing, cosmetics) and age.

Unfortunately, there is yet no public database available that contains facial images which exhibit all of the mentioned properties. Therefore, we decided to set up a new MAD database based on existing face image databases. In this section, the required steps, including the selection of the images from public face image databases, for the generation of the morphed images, the pairing of reference and probe images, and the application of post-processings to the images, are described in detail.

A. Selection of the Facial Image Database

In a first step, we selected public databases from which we could build our MAD database. The candidate database had to fulfill the following conditions:

- It must contain images (meeting the subsequent conditions) of a sufficient amount of different subjects.
- The images must have a sufficient resolution.
- For each subject, at least two ICAO-compliant face images must be available (one of which used as input for morph generation and the other as bona fide).
- For each subject, at least one image taken in less constrained conditions (resembling the border control scenario) must be available.

Among the available databases, FERET [12] and FRGCv2 [8] are suitable. The samples of FERET are all taken in a controlled environment but contain variations in pose and expression. FRGCv2 contains images suitable as passport photos, but also images with scenario variations, e.g., non-uniform illumination, lack of sharpness and uneven background, suitable as probe images.

The FERET and the FRGCv2 face databases were collected over a period of three and two years, respectively. This means the maximum age variation of captured subjects is limited by these periods. That is, selected face images of the used databases exhibit less variations with respect to temporal difference between the capturing of reference and probe images than those expected in a border control scenario. However, to the best of the authors' knowledge, publicly available databases which contain face images captured over a larger period of time do not fulfill the aforementioned requirements. Moreover, we stress that large age variations are only expected to occur for bona fide authentication attempts. For an MA it is more plausible that it will be performed using current face images of the attacker and the accomplice which results in a small temporal difference between capturing of reference and probe images.

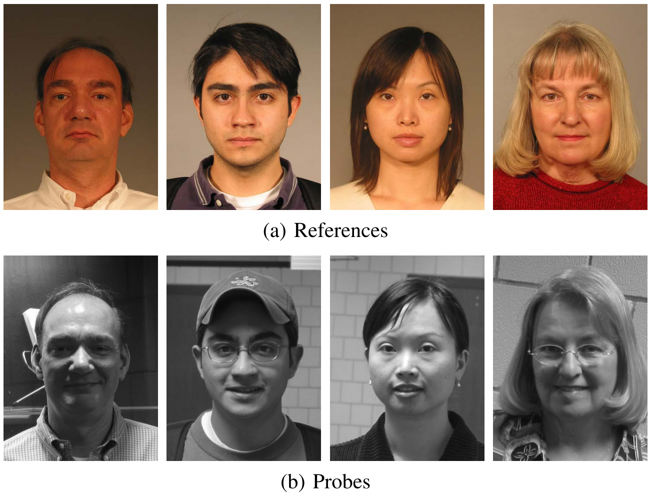
B. Selection of Image Candidates

For each of these databases, we select two types of images, representing the reference, i.e., passport photographs, and the unconstrained probe images, i.e., trusted live captures.

The set of reference images contains all images complying with ICAO requirements [46] except alignment of the face. These images exhibit, among other requirements, uniform illumination, good focus, a neutral face expression with open eyes and no visible teeth, neutral background and no reflections in glasses. For these images, we adjusted the alignment of the face by suitable scaling, rotation and padding/cropping to ensure that the ICAO requirements with respect to the eyes' positions are met. Precisely, facial landmarks are detected applying the Dlib algorithm [47] and alignment is performed with respect to the detected eye coordinates where a fixed intra-eye distance of 180 pixels is ensured.

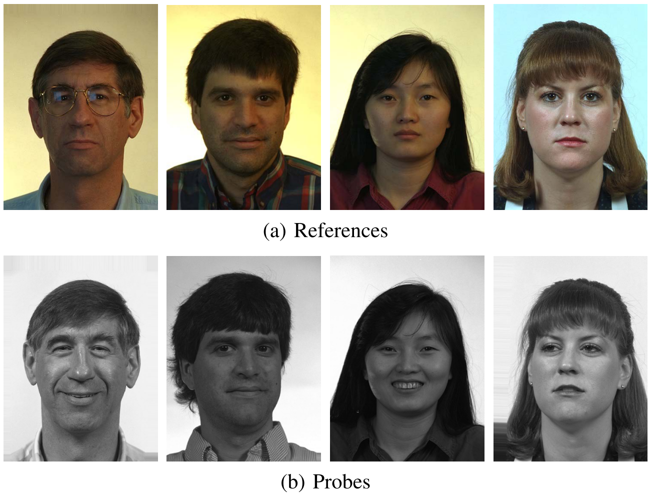
From the remaining images, those suitable as probe images to resemble the border crossing capture process were selected. The face should be recognizable, but can be only partially illuminated and slightly out of focus. The selected images of the FERET database yield a variance in facial expression and pose (slight rotations). For FRGCv2, the selected images yield a variance in facial expression, background, illumination and sharpness. The probe images are additionally converted to grayscale. This conversion is motivated by the fact that some eGates capture grayscale images. Nonetheless, if face image quality is sufficient, the use of grayscale probe images is expected to have negligible impact on the recognition accuracy of state-of-the-art face recognition systems as well as human examiners [48], [49].

Examples of the reference and probe images of both databases are shown in Figure 2 and Figure 3. Note that the criteria used to select suitable image candidates, which are necessary to resemble a realistic scenario, clearly limit the amount of appropriated images. Eventually, it is important to note that the selection of face images is to a certain degree subjective. In order to facilitate reproducibility of this research the list of selected face images is made available to the research community upon request.


(a) References

(b) Probes

Fig. 2. Examples of reference and gray scale probe images for FRGCv2.


(a) References

(b) Probes

Fig. 3. Examples of reference and gray scale probe images for FERET.

C. Selection of Images and Morph Creation

For each subject, we selected from all available reference images a subset of *bona fide reference images* and a subset of images for generating morphs, i.e., *morph input images*, as well as a subset of *probe images*. Where possible, we ensured that the bona fide reference images were disjoint from the morph input images, thereby avoiding a repetitive use of the same image or same image parts during MAD training which increases the variance of training data. No lossy compression was applied to the images prior to post-processing. In addition, we assured that selected reference images (bona fide or morph input images), if possible, are captured in a different recording session than the probe images, in order to achieve temporal difference between the capturing of reference and probe images as expected in a real border control scenario. For image candidates selected from the FERET database, where possible, we selected one bona fide reference image, one morph input image, and up to two probe images per subject. For image candidates selected from the FRGCv2 database, if possible, we chose two bona fide reference images, two morph input images, and up to five probe images per subject.

TABLE II

COMPOSITION OF THE MAD DATABASE. THE NUMBER OF MORPH IMAGES IS MULTIPLIED BY THE NUMBER OF MORPHING TOOLS USED

Database	Subjects	Male	Female	Reference Images		Probes
				Bona Fide	Morphs	
FERET	529	329	200	529	529*	791
FRGCv2	533	302	231	984	964*	1,726

For the creation of morphs, the morph input images are lexicographically sorted, and then each input image is morphed with one of the next consecutive input image, for which the following two criteria hold: both depicted subjects are of same sex; only one face image of the image pair shows glasses. The latter criteria should avoid obvious artefacts. Morph input images are only used once, i.e., for the generation of a single morph. It is important to note that after the morph generation the morph input images are discarded. Again, we aim to avoid a repeated use of the same image or same image parts in the training stage of MAD algorithms which should also prevent from overfitting caused by overrepresented image parts. The list of resulting face image pairs used for morph generation is provided to the research community upon request. The key figures of the resulting database are listed in Table II.

D. Morphing Algorithms

Four different automated morphing tools were used to create morphed face images:

1) **FaceFusion**,[2] a proprietary morphing algorithm. Originally being an iOS app, we deployed an adaptation for Windows which uses the 68 landmarks of Dlib and Delaunay triangles. After the morphing process, certain regions (eyes, nostrils, hair) of the first face image are blended over the morph to hide artefacts. The corresponding landmarks of upper and lower lips can be reduced as described in [17] to avoid artefacts at closed mouths. The created morphs have a high quality and low to no visible artefacts. Example morphs are shown in Figure 4b.
2) **FaceMorpher**,[3] an open-source implementation using Python. In the version applied for this work, the algorithm uses STASM for landmark detection. Delaunay triangles, which are formed from the landmarks, are wrapped and blended. The area outside the landmarks is averaged. The generated morphs show strong artefacts in particular in the area of neck and hair. Figure 4c depicts examples of generated morphs.
3) **OpenCV**, a self-made morphing algorithm derived from "Face Morph Using OpenCV".[4] This algorithm works similar to FaceMorpher. Important differences between the algorithms are that for landmark detection Dlib is used instead of STASM and that additional landmarks are positioned at the edges of the image, which are also used to create the morphs. Thus, in contrast to FaceMorpher, the outer facial area does not consist of an averaged image, but like the rest of the image, of morphed triangles. However, visible artefacts outside the face area can be observed, which is mainly due to missing landmarks. Example morphs can be seen in Figure 4d.
4) **UBO-Morpher**, the morphing tool of University of Bologna, as used, e.g., in [13]. Dlib landmarks were used for this algorithm. The morphs are generated by triangulation, averaging and blending. To avoid the artefacts in the area outside the face, the morphed face is copied to the background of one of the original images. Even if the colors are adjusted at boundaries, visible edges may appear at the transitions. Figure 4e shows examples of resulting morphs.

No manual post-processing is applied in the morph generation process. The use of various morphing tools which produce facial morphs of different quality enable a thorough investigation of MAD capabilities, i.e., impact of the use of high and low quality morphs on training and testing of MAD methods, see Section V. The evaluation of MAD algorithms for different quality levels of morphed face images is also performed by NIST in the FRVT MORPH [22]. The pairs for the morphing process for each algorithm are selected according the protocol defined in Section III-C, resulting in 4 × 529 morphed face images for FERET and 4 × 964 morphed face images for FRGCv2. For both databases, the bona fide and morphed face images are normalized to meet the ICAO-requirements for passport images [46]. The resulting images are of size 720 × 960 pixels.

E. Post-Processing

Images that have been captured for the use in an identity document, e.g., passport, can go through various processing steps before they are embedded, e.g., in a passport RFID chip. To reflect this variety, the passport images (bona fide and morphed) of the MAD database are post-processed in different manners. An example for the different post-processings is shown in Figure 5.

Unprocessed: The images are not further processed, henceforth referred to as *NPP* (no post-processing). NPP images serves as baseline.

Resized: The resolution of the images is reduced by half, resulting in 360 × 480 pixels, in the following text referred to as *Resized*. Resized images fulfill the minimum requirement with respect to intra-eye distance defined by ICAO, i.e., 90 pixels. This pre-processing corresponds to the scenario that an image is submitted digitally by the applicant.

JPEG2000: The images are resized by half and then compressed using JPEG2000, a wavelet-based image compression method that is recommended for EU passports [50]. The setting is selected in a way that a target file size of 15KB is achieved. This scenario reflects the post-processing path of passport images if handed over digitally at the application desk, hereafter referred to as *JP2*.

[2] www.wearemoment.com/FaceFusion/

[3] github.com/alyssaq/face_morpher

[4] www.learnopencv.com/face-morph-using-opencv-cpp-python/

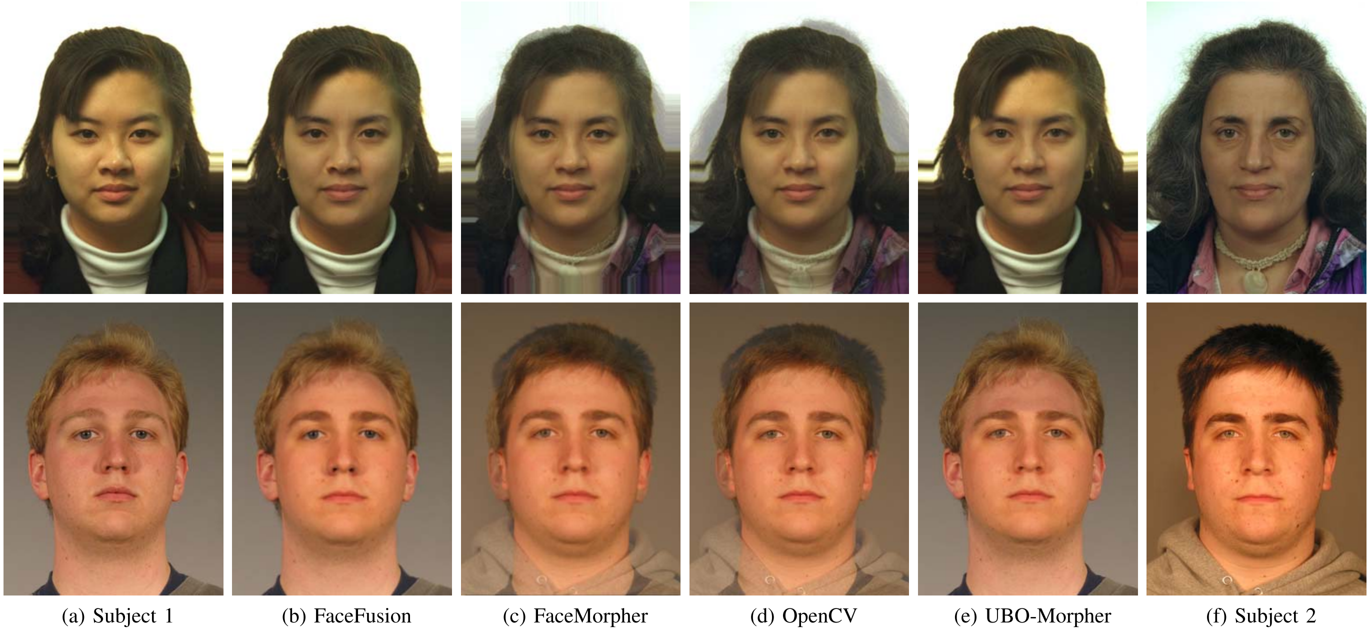


Fig. 4. Examples for a morphed face images from all four algorithms for a female pair of face images from the FERET (top row) and the FRGCv2 (bottom row). Equal weights of face images were used to create morphs. Face images depicting subjects with similar demographic attributes, e.g., age and ethnicity, generally result in more plausible face morphs.

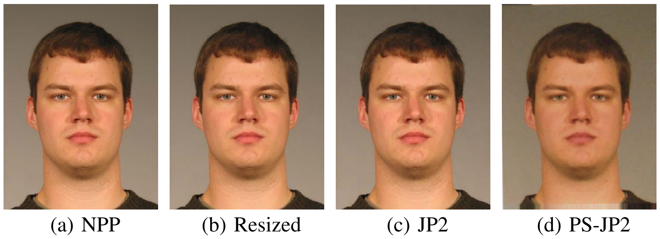


Fig. 5. Examples of an original image and the three post-processing types.

Print/Scan – JPEG2000: The original images (uncompressed and not resized) are first printed with a high quality laser printer (*Fujifilm Frontier 5700R Minlab* on *Fujicolor Crystal Archive Paper Supreme HD Lustre photo paper*) and then scanned with a premium flatbed scanner (*Epson DS-50000*) with 300 dpi. A dust and scratch filter is then applied in order to reduce image noise. Subsequently, the images are resized to 360 × 480 pixels, i.e., half of the NPP images, and compressed to 15 KB using JPEG2000.[5] This scenario reflects the post-processing path of passport images if handed over at the application desk as a printed photograph, subsequently referred to as *PS-JP2*.

F. Validation of Attack Potential

The newly created databases differ considerably from other databases used in scientific publications on MA and MAD. In particular, the intra-class variation is much higher in our database due to the selection of unconstrained probe images. While this approach ensures that our database is more eligible to simulate real-world scenarios, it is perfectly valid to question whether the use of unconstrained probe images may render MA ineffective. Previous analyses of the vulnerability of face recognition systems to MAs, e.g., in [4], [7], used probe images, which nearly resembled passport images and, hence, these studies do not apply to the face databases used in this work. Therefore, we evaluated whether face recognition systems are also vulnerable to MAs using our newly created databases.

TABLE III
NUMBER OF COMPARISONS PER TEST SET

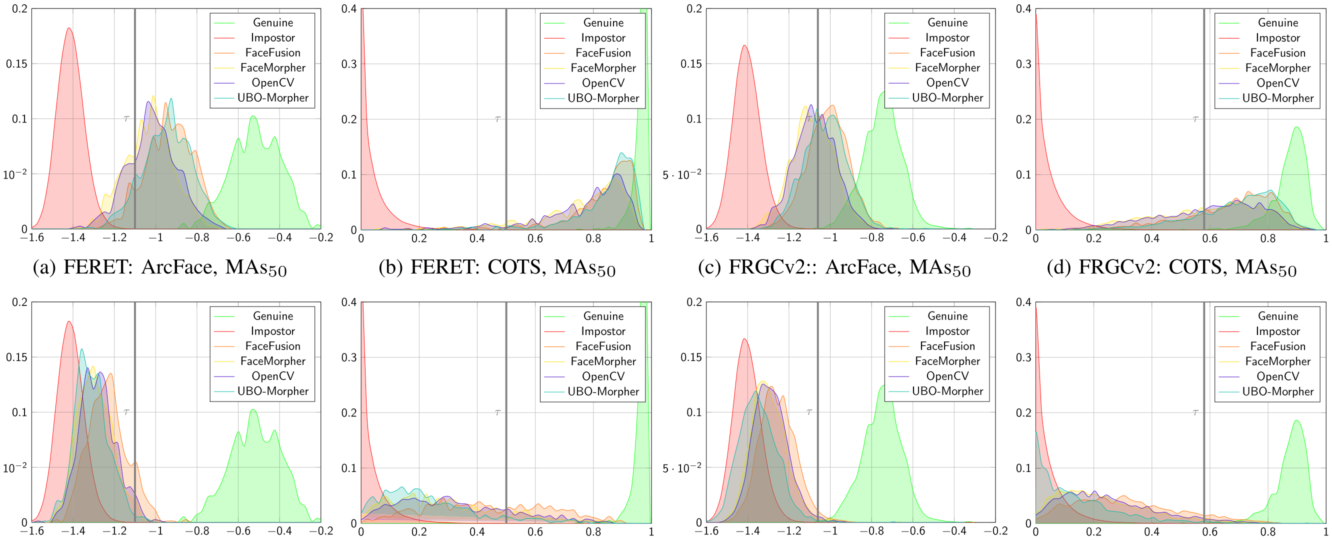
Database	Bona Fide	Morph Attacks	Impostors
FERET	791	791	418,966
FRGCv2	3,298	3,246	1,695,086

Firstly, the face recognition performance is evaluated. The amount of bona fide, i.e., genuine, and impostor comparisons, i.e., all possible cross-comparisons of reference and probe images of different subjects, for each database is summarized in Table III. On each database, the decision thresholds of the face recognition systems are set to achieve a False Match Rate (FMR) of 0.1% according to the FRONTEX recommendation for border control scenarios [51]. The results are presented in Table IV and the corresponding Probability Density Functions (PDFs) are depicted in Figure 6. For ArcFace and the COTS system, no false non-matches occur at an FMR of 0.1%. Note that this is also to be expected for using color probe images. In general, it can be observed, that the genuine score distributions of FERET are further right (higher similarity scores) than those of FRGCv2. This is likely due to

[5] Due to the lustre print, the scans exhibit a visible pattern of the paper surface, which is only partly removed by the dust and scratch filter and results in stronger compression artefacts than for scans of glossy prints.

TABLE IV
VULNERABILITY ASSESSMENT OF FACE RECOGNITION SYSTEMS (FRSs)

FRS	Database	Decision Threshold	FNMR at FMR=0.1%	MMPMR/RMMR (in %)							
				FaceFusion		FaceMorpher		OpenCV		UBO-Morpher	
				MAs_{50}	MAs_{25}	MAs_{50}	MAs_{25}	MAs_{50}	MAs_{25}	MAs_{50}	MAs_{25}
ArcFace	FERET	-1.10	0.0	93.0	17.3	75.7	5.0	79.9	6.7	92.9	2.3
	FRGCv2	-1.06	0.0	77.2	14.7	50.0	8.9	54.0	9.2	72.4	7.2
COTS	FERET	0.54	0.0	96.0	33.1	88.6	10.0	91.3	13.0	96.9	5.5
	FRGCv2	0.58	0.0	79.4	18.8	60.1	8.4	62.8	9.8	81.5	3.0



(a) FERET: ArcFace, MAs_{50} (b) FERET: COTS, MAs_{50} (c) FRGCv2:: ArcFace, MAs_{50} (d) FRGCv2: COTS, MAs_{50}

(e) FERET: ArcFace, MAs_{25} (f) FERET: COTS, MAs_{25} (g) FRGCv2: ArcFace, MAs_{25} (h) FRGCv2: COTS, MAs_{25}

Fig. 6. Probability Density Functions of comparison scores of genuine, impostor and morph comparisons for all three face recognition systems. Distance scores obtained from ArcFace were multiplied by -1 to obtain similarity scores. τ depicts the estimated threshold for a FMR of 0.1%.

the fact, that the probes of FRGCv2 contain a much higher variation in illumination, sharpness and expression. Further, it can be seen, that the impostor distributions of FERET and FRGCv2 are close to each other, thus the thresholds are approximately the same over both datasets.

Secondly, the vulnerability of the face recognition systems against MAs is evaluated. The vulnerability assessment was conducted using the metrics presented described in [28], i.e., Mated Morph Presentation Match Rate (MMPMR) and Relative Morph Match Rate (RMMR). The MMPMR describes the proportion of morphed face images accepted by the face recognition system, the RMMR describes the relation between the MMPMR and the true match rate. Since on both databases no false non-matches occur for the considered decision threshold the MMPMR and the RMMR are equal. The vulnerability analysis is performed exclusively on the NPP image sets of each database, as preliminary studies have shown that the performance of facial recognition systems is only slightly affected by the previously described post-processings. An open-source face recognition system, namely ArcFace [52], and one commercial off-the-shelf system, referred to as COTS,[6] are employed in the vulnerability assessment. The number of MAs performed for each database is summarized in Table III. The corresponding results are depicted in Figure 6 and summarized in Table IV.

Two types of MAs are considered in the vulnerability analysis. On the one hand, morphs with equal weights are employed, i.e., facial images of both subjects contribute with a weight of 50% to the resulting morph. MAs based on these morphs are denoted as MAs_{50}. On the other hand, morphs in which the attacker contributes only with 25% to the morph are used. Corresponding MAs are referred to as MAs_{25}. This means, a larger weight is assigned to an accomplice, i.e., 75%, who would present the morphed face image to a human observer during the application process of an identity document. Assuming that morphs generated with equal weights of two face images are easily spotted by a human inspection during the application process, the latter type of MAs might be considered as more realistic [13].

Focusing on MAs_{50}, ArcFace and the COTS system are both very vulnerable to the MAs contained in the database. This can be observed from the high MMPMR/RMMR values in Table IV. Especially on the FERET dataset almost all MAs are accepted. In contrast, the MMPMR/RMMR values, i.e., success chance, significantly drop in case of MAs_{25}. In general, morphs created by morphing algorithms producing less artefacts (FaceFusion and UBO-Morpher) are more likely

[6] We stress that this COTS system is not Eyedea, which is only used for MAD.

to pass the recognition system. In addition, it should be noted that the morphs generated from the FERET database are generally more successful in MAs than the morphs from FRGCv2. This can be attributed to the different intra-class variations of the two databases. Since the genuine comparisons of the FERET data set achieve higher similarity scores, the ones obtained by MAs also tend to be higher. As MAs$_{25}$ achieve only small MMPMR/RMMR values only MAs$_{50}$ are considered for MAD. In case the attacker contributes with a lower weight to the morph compared the accomplice the distance between the resulting morph and the probe image of the attacker increases. Consequently, differential MAD method will generally detect such MAs with higher chance, compared to MAs where morphs have been created using equal weights of the contributing face images.

IV. MAD Based on Deep Face Representations

One of the greatest issue regarding existing differential MAD algorithms is that they can not cope with the large intra-class variance that must be expected for realistic probe images. Deep face recognition networks, however, have shown that they are able to work very robustly, even on challenging data. Therefore, we propose to employ deep face representations extracted by such deep face recognition systems for differential MAD. Precisely, we use the following deep face recognition systems:

- ArcFace [52], an up-to-date network with a topology optimized for automatic facial recognition
- The commercial face recognition SDK from Eyedea[7]
- A re-implementation[8] of the FaceNet algorithm [53]

In principle, it would by possible to apply transfer learning and re-train a pre-trained deep face recognition network to detect morphs. However, the high complexity of the model, represented by the large number of weights in the neural network, requires a large amount of training data. Even if only the lower layers are re-trained, as done in [18], the limited number of training images (and much lower number of subjects) in our database can easily result in overfitting to the characteristics of the training set.

Therefore, we follow an alternative approach. We use the pre-trained deep face recognition networks as feature extractors and train our MAD algorithms on the deep representations extracted by the neural network (on the lowest layer). Deep face recognition systems leverage very large databases of face images to learn rich and compact representations of faces. While these feature vectors, i.e., deep face representations, have not been trained to detect MAs, at least in the differential scenario, they can, nevertheless, be very useful for MAD: As a morphed face image does not only contain biometric information of the attacker but also those of the accomplice, its deep face representation is expected to, at least in certain aspects, considerably deviate from those detected in the probe image. On the other hand, since there were no morphed facial images in the training set of the neural network, the features can not contain information on image characteristics specific for a certain morphing technique or tool, which reduces the risk of overfitting.

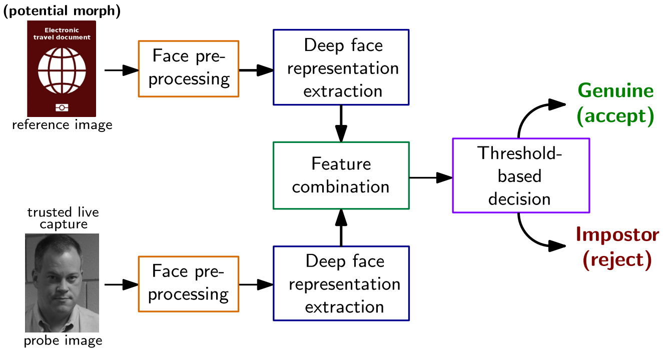


Fig. 7. Generic processing chain of face recognition systems based on neural networks.

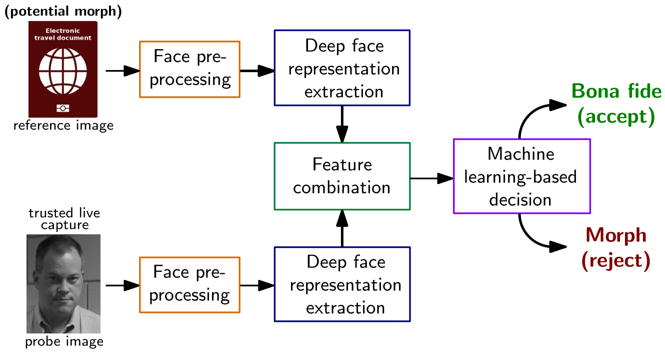


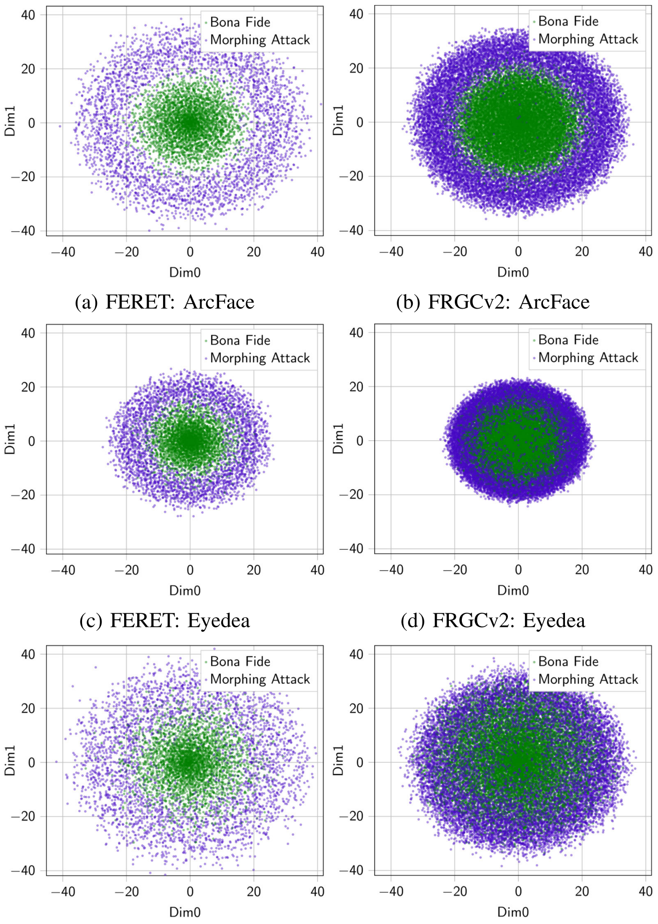
Fig. 8. Overview of the proposed differential MAD system based on deep face representations.

Most deep face recognition systems work as shown in Figure 7. The facial image is pre-processed and transferred to the neural network trained for the extraction of deep face representations. This net transfers the facial image into a discriminatory feature space with smaller dimension (512 in the case of ArcFace and FaceNet, 256 in the case of Eyedea). If two images are to be compared, the distance of their feature vectors, e.g., using the Euclidean distance, can serve as dissimilarity score. Even though our vulnerability analysis in Section III-F has shown that this measure is not suitable to separate morphs from bona fide images, the feature vectors can nevertheless contain sufficient information to detect MAs.

The processing of our MAD algorithms is shown in Figure 8. In the training stage, pre-processed bona fide or morphed reference images and probe images are fed into the neural network. The resulting deep face representations are combined and subsequently processed by a machine learning-based classifier to learn to distinguish between bona fide authentication attempts and MAs. During testing, a potentially morphed reference image and a probe image are processed in the same way and the previously trained classifier is used to estimate the MAD score. A simple but effective combination of the deep features is subtraction, i.e. estimation of a difference vector, which preserves the dimensionality

[7] https://www.eyedea.cz/eyeface-sdk/

[8] David Sandberg - Face Recognition using Tensorflow, URL: https://github.com/davidsandberg/facenet



(a) FERET: ArcFace (b) FRGCv2: ArcFace

(c) FERET: Eyedea (d) FRGCv2: Eyedea

(e) FERET: FaceNet (f) FRGCv2: FaceNet

Fig. 9. Scatter plots of MDS-reduced difference vectors of deep face representations extracted from reference and probe images.

of the feature vector and keeps the training effort low. For the classifier, we tested various machine learning algorithms, namely AdaBoost, Random Forest, Gradient Boosting and SVM. Consistently, SVM with radial basis function as kernel showed the best accuracy and was, thus, chosen for the MAD algorithm.

By employing the difference vectors of feature vectors extracted from reference and probe images, a simple dimension reduction by Multi-Dimensional Scaling (MDS) to two dimensions shows that bona fide and morphs can be separated. Corresponding scatter plots are provided in Figure 9. Each plot shows data points for MAs using all of the four morphing tools. Since, as shown in Section III-C, the subset of the FRGCv2 allows more comparisons, the corresponding scatter plots are much denser compared to the ones generated from the FERET database. Nevertheless, for both databases it can be seen that MAs and bona fide authentication attempts can be separated on the basis of the difference vectors of deep face representations obtained from ArcFace and Eyedea. Especially the FERET database can be separated almost error-free using the ArcFace algorithm for feature extraction. The FaceNet feature vectors also allow a separation, but higher error rates are expected.

V. Experiments

Using the databases described in Section III, we evaluate our MAD approach based on deep face representations described in Section IV. In our first experiment, we benchmark the accuracy of our approach together with that of other differential MAD algorithms in the absence of image post-processing. Then, we evaluate the robustness of our approach against image post-processing. Finally, we analyze the score distributions in more detail to gain insight on the causes of classification errors.

For all evaluations presented in this section, we separate training and test sets by source database; precisely, all algorithms are trained using the default hyperparameters on images originating from FERET and evaluated on images from FRGCv2, and vice versa. This approach does not only ensure a strict separation of training and test data, but also a large variance in the image characteristics between these sets. Furthermore, for all evaluations, each of the training sets contains only images with the one of the four post-processings and morphs created with one of the four morphing algorithms (see Section III for details); i.e., we do not combine several post-processings or morphing tools in one training set. The numbers of MAs and bona fide authentication attempts performed for each database during training and testing are listed in Table III.

The accuracy of the detection algorithms is reported using the Detection Equal Error Rate (D-EER), i.e., at the decision threshold where the proportion of attack presentations incorrectly classified as bona fide presentations (APCER) is as high as the proportion of bona fide presentations incorrectly classified as presentation attack (BPCER). For APCER and BPCER the definitions of ISO IEC 30107-3 [54] for measuring accuracy of presentation attack detection are used:

APCER: proportion of attack presentations incorrectly classified as bona fide presentations in a specific scenario

BPCER: proportion of bona fide presentations incorrectly classified as presentation attacks in a specific scenario

Additionally, the BPCER10 is reported, i.e., the operation point where APCER $= 10\%$.

A. Detection Accuracy in the Absence of Post-Processing and Comparison With Other MAD Algorithms

In this subsection, we evaluate the accuracy of our MAD approach on the NPP images and compare its detection accuracy to that of several other differential MAD approaches. Beside the classifiers trained on deep face representations extracted using ArcFace, Eyedea and FaceNet, we test two MAD algorithms based on texture descriptors, namely LBP [26] and BSIF [27] with patches of size 3×3 and an optional division into 4×4 cells. In order to adapt these feature extractors to differential detection scenario, the difference between the histograms of reference and probe images are used as classifier input as proposed in [7]. Further, two landmark-based algorithms implemented according to [11] are

TABLE V

Detection Performance of MAD Algorithms on NPP Images Using Different Morphing Tools (MTs) for Training and Test

			D-EER | BPCER-10 (in %)									
Train DB	**Train MT**	**Test MT**	ArcFace	FaceNet	Eyedea	BSIF	BSIF$_{4\times4}$	LBP	LBP$_{4\times4}$	LM-Wing	LM-Dlib	Demorphing
FERET	FaceFusion	FaceFusion	**6.2** | **3.5**	25.9 | 52.7	15.1 | 22.6	39.6 | 75.8	39.8 | 77.5	38.7 | 82.4	41.0 | 82.0	42.5 | 83.3	44.5 | 83.0	8.6 | 4.4
		FaceMorpher	3.6 | 0.7	21.7 | 41.0	12.4 | 14.7	41.8 | 83.5	43.5 | 86.5	38.6 | 91.6	42.7 | 92.7	41.5 | 81.2	43.6 | 83.0	**3.4** | **0.0**
		OpenCV	**3.7** | **0.8**	21.7 | 40.7	12.3 | 14.8	42.8 | 82.8	43.5 | 83.7	39.7 | 86.5	42.6 | 87.9	42.5 | 82.1	43.9 | 82.8	3.8 | **0.8**
		UBO-Morpher	**6.0** | **3.5**	25.5 | 50.5	16.2 | 24.5	42.5 | 81.4	43.3 | 83.0	39.2 | 85.1	42.8 | 85.3	42.3 | 82.3	44.4 | 83.3	7.8 | 5.8
	FaceMorpher	FaceFusion	**7.1** | 4.5	26.3 | 53.4	16.2 | 24.9	49.3 | 83.3	46.7 | 83.4	48.1 | 87.1	46.3 | 87.5	43.9 | 81.7	43.8 | 88.1	8.6 | **4.4**
		FaceMorpher	**3.3** | 0.5	19.8 | 36.4	10.3 | 10.4	50.2 | 88.0	47.5 | 88.6	48.1 | 93.2	45.5 | 95.2	42.0 | 79.1	43.0 | 87.7	3.4 | **0.0**
		OpenCV	**3.7** | **0.7**	20.7 | 37.8	10.5 | 11.5	51.3 | 86.5	49.1 | 86.5	48.8 | 89.0	47.1 | 90.7	43.2 | 80.0	42.7 | 87.2	3.8 | 0.8
		UBO-Morpher	**6.6** | **4.2**	25.2 | 49.5	17.2 | 26.3	51.1 | 86.5	48.7 | 85.3	48.1 | 88.8	47.4 | 89.1	43.7 | 81.6	42.9 | 87.1	7.8 | 5.8
	OpenCV	FaceFusion	**6.8** | **4.0**	26.8 | 53.6	16.3 | 23.9	46.9 | 82.0	43.3 | 80.7	44.9 | 86.0	43.7 | 84.8	43.1 | 83.2	45.5 | 84.2	8.6 | 4.4
		FaceMorpher	**3.2** | 0.7	20.1 | 36.2	11.0 | 12.0	46.4 | 84.9	44.9 | 87.1	42.8 | 92.9	43.2 | 94.8	41.2 | 80.9	44.2 | 84.0	3.4 | **0.0**
		OpenCV	**3.3** | **0.8**	20.1 | 35.8	10.7 | 11.9	48.2 | 84.7	45.8 | 85.2	44.9 | 88.9	44.3 | 88.9	42.0 | 81.5	44.4 | 83.7	3.8 | **0.8**
		UBO	**6.3** | **3.6**	24.9 | 49.9	16.8 | 24.2	48.3 | 84.2	45.3 | 84.0	44.8 | 88.5	44.8 | 87.1	42.8 | 83.0	45.1 | 84.6	7.8 | 5.8
	UBO-Morpher	FaceFusion	**6.7** | **4.2**	26.6 | 54.2	16.4 | 24.2	47.0 | 81.0	44.4 | 81.7	43.7 | 84.2	44.0 | 85.8	42.7 | 82.5	44.8 | 84.3	8.6 | 4.4
		FaceMorpher	3.8 | 1.1	21.8 | 42.3	12.8 | 17.1	45.7 | 84.7	46.7 | 87.9	39.7 | 91.2	44.2 | 94.1	41.5 | 80.9	44.2 | 86.0	**3.4** | **0.0**
		OpenCV	**3.8** | 1.2	21.6 | 39.3	12.7 | 16.5	47.1 | 84.5	46.7 | 85.4	42.6 | 87.1	44.2 | 89.5	42.5 | 81.5	44.4 | 84.2	**3.8** | **0.8**
		UBO-Morpher	**5.7** | **3.0**	24.4 | 48.7	15.5 | 22.7	47.2 | 82.8	45.6 | 83.5	41.6 | 85.4	43.0 | 85.6	41.6 | 82.0	44.3 | 82.4	7.8 | 5.8
FRGCv2	FaceFusion	FaceFusion	**2.1** | **0.6**	13.0 | 15.4	5.3 | 3.1	15.5 | 23.8	15.9 | 28.3	18.1 | 30.2	25.9 | 49.1	39.4 | 84.4	38.9 | 75.7	6.7 | 5.1
		FaceMorpher	**1.1** | **0.0**	9.6 | 9.4	4.3 | 1.9	14.5 | 18.8	13.8 | 24.6	12.9 | 17.4	21.9 | 38.1	41.2 | 85.5	37.6 | 71.7	4.1 | 1.3
		OpenCV	**1.4** | **0.0**	9.8 | 9.4	4.0 | 2.4	15.9 | 27.0	16.7 | 30.2	16.9 | 28.2	24.3 | 47.9	41.3 | 87.0	39.4 | 75.8	4.2 | 1.4
		UBO-Morpher	**2.5** | **0.9**	13.0 | 15.4	6.9 | 5.1	16.3 | 25.5	18.1 | 32.4	17.0 | 27.2	24.3 | 45.3	40.2 | 86.2	39.5 | 77.5	7.7 | 6.1
	FaceMorpher	FaceFusion	**2.7** | **1.1**	14.4 | 18.1	7.3 | 5.7	24.9 | 47.5	23.1 | 52.0	33.0 | 63.9	37.7 | 75.7	40.6 | 82.5	39.8 | 78.6	6.7 | 5.1
		FaceMorpher	**0.9** | **0.0**	8.4 | 6.5	3.1 | 1.2	15.1 | 21.7	15.5 | 24.3	24.6 | 46.2	31.7 | 63.1	40.8 | 83.5	39.8 | 83.7	4.1 | 1.3
		OpenCV	**1.1** | **0.0**	8.2 | 6.8	3.9 | 1.6	19.3 | 33.0	19.3 | 40.1	30.2 | 57.2	34.6 | 69.6	41.9 | 84.8	40.8 | 81.9	4.2 | 1.4
		UBO-Morpher	**2.5** | **1.1**	13.6 | 16.0	8.3 | 7.3	21.6 | 36.4	20.7 | 44.0	30.3 | 57.5	35.4 | 70.3	41.0 | 83.2	42.1 | 87.0	7.7 | 6.1
	OpenCV	FaceFusion	**2.7** | **0.9**	14.0 | 19.2	7.3 | 5.3	18.4 | 31.9	19.3 | 34.0	23.9 | 44.8	30.7 | 62.4	40.8 | 83.2	38.2 | 75.2	6.7 | 5.1
		FaceMorpher	**0.6** | **0.0**	9.3 | 8.4	2.9 | 1.1	14.1 | 18.3	15.3 | 23.9	18.6 | 33.4	25.5 | 51.3	41.3 | 85.9	35.6 | 73.1	4.1 | 1.3
		OpenCV	**0.9** | **0.0**	8.9 | 7.9	3.4 | 1.2	15.4 | 25.3	18.1 | 30.6	21.9 | 43.0	28.8 | 58.7	41.2 | 86.3	38.0 | 75.3	4.2 | 1.4
		UBO-Morpher	**2.7** | **0.9**	13.1 | 16.9	7.5 | 6.7	17.4 | 26.5	18.1 | 32.4	21.7 | 42.7	29.3 | 57.0	41.9 | 85.8	38.8 | 77.7	7.7 | 6.1
	UBO-Morpher	FaceFusion	**3.2** | **0.7**	13.4 | 16.7	6.0 | 4.0	20.5 | 33.9	18.8 | 35.0	21.7 | 46.7	31.0 | 61.0	41.1 | 86.3	38.4 | 76.6	6.7 | 5.1
		FaceMorpher	**0.7** | **0.0**	9.8 | 9.6	4.4 | 2.4	13.6 | 19.6	14.5 | 22.2	16.8 | 25.5	25.4 | 49.7	43.0 | 86.7	36.0 | 71.3	4.1 | 1.3
		OpenCV	**1.1** | **0.0**	9.4 | 9.1	4.1 | 2.4	16.4 | 24.3	18.2 | 28.3	20.2 | 41.6	28.8 | 58.6	42.5 | 88.4	39.4 | 74.7	4.2 | 1.4
		UBO-Morpher	**2.4** | **0.7**	11.5 | 13.6	5.4 | 4.0	17.5 | 23.9	18.2 | 27.8	19.8 | 38.8	28.4 | 56.5	40.8 | 87.2	38.7 | 75.5	7.7 | 6.1

evaluated using landmarks computed with Dlib [47] and Wing Loss [55], respectively. Finally, we evaluate the demorphing algorithm of [13] with a demorphing factor of 0.3 in combination with Dlib landmarks and the COTS face recognition algorithm used in Section III-F.

The results are shown in Table V. Note that the demorphing algorithm is not trained, which is why for this MAD algorithm in the table the specification of the morphing algorithm used for training have no effect. Consistently, when testing on FRGCv2 (and thus training on FERET), the detection performance of all MAD algorithms is much lower than that achieved on FERET. This observation can be explained by the fact that, as visible in Figure 2, the probes of the FRGCv2 show a significantly higher variance in illumination, background and sharpness, whereas the probes of the FERET database contain pose variants, but are consistently good in quality, see Figure 3. Similarly, our results on the MAD based on texture descriptors and landmarks are much worse than those reported in [7] and [11], where probe images of higher quality were used. Thus, we conclude that the quality of the probes has a strong effect on the detection performance of the MAD algorithms. This further underlines the need for databases containing probe images with realistic characteristics.

Our MAD algorithms based on deep face representations yields superior results compared to the majority of other methods. The algorithm based on ArcFace features by far outperforms all other approaches, achieving very low D-EER between 1% and 7%. The algorithm using Eyedea features ranks second with an D-EER between 3% and 17%. Obviously, the features of FaceNet are far less suitable for MAD. A clear correlation to the scatter plots shown in Figure 9 can be observed.

The texture-based MAD algorithms are only capable of detecting morphed face images (although with very high error rates) if the probe image is available in sufficient quality, i.e., on the FERET database. The detection performance of these algorithms when probes with a more realistic variance, i.e., on the FRGCv2 database, is close to random. The performance of landmark-based MAD is less dependent on the quality of probe images but generally very poor. In contrast, demorphing in combination with the COTS face recognition system yields good detection rates. In particular, for FRGCv2, it performs only slightly worse than our ArcFace-based MAD

TABLE VI

DETECTION PERFORMANCE OF MAD ALGORITHMS BASED ON DEEP FACE REPRESENTATIONS WITH DIFFERENT POST-PROCESSINGS (PPS)

<table>
<tr><th colspan="4"></th><th colspan="12">D-EER | BPCER-10 (in %)</th></tr>
<tr><th colspan="4"></th><th>NPP</th><th>Resized</th><th>JP2</th><th>PS-JP2</th><th>NPP</th><th>Resized</th><th>JP2</th><th>PS-JP2</th><th>NPP</th><th>Resized</th><th>JP2</th><th>PS-JP2</th></tr>
<tr><th>Train DB</th><th>Train MA</th><th>Test MA</th><th>Test PP</th><th colspan="4">ArcFace</th><th colspan="4">Eyedea</th><th colspan="4">FaceNet</th></tr>
<tr><td rowspan="16">FERET</td><td rowspan="8">FaceMorpher</td><td rowspan="4">FaceFusion</td><td>NPP</td><td>7.3 | 4.6</td><td>7.2 | 4.6</td><td>7.1 | 5.0</td><td>7.3 | 5.3</td><td>16.2 | 24.6</td><td>16.2 | 24.9</td><td>16.1 | 25.3</td><td>16.5 | 25.5</td><td>27.5 | 54.5</td><td>27.5 | 54.3</td><td>27.5 | 55.4</td><td>28.1 | 56.8</td></tr>
<tr><td>Resized</td><td>7.1 | 4.5</td><td>7.1 | 4.6</td><td>7.1 | 5.0</td><td>7.3 | 5.1</td><td>16.2 | 24.9</td><td>16.2 | 25.1</td><td>16.3 | 24.9</td><td>16.7 | 25.8</td><td>26.3 | 53.4</td><td>25.8 | 53.4</td><td>26.3 | 52.6</td><td>26.7 | 56.3</td></tr>
<tr><td>JP2</td><td>7.7 | 5.9</td><td>7.7 | 6.0</td><td>7.4 | 6.2</td><td>7.7 | 6.2</td><td>16.8 | 26.7</td><td>16.9 | 26.1</td><td>16.5 | 26.1</td><td>16.8 | 26.0</td><td>26.9 | 54.1</td><td>26.8 | 54.2</td><td>26.8 | 53.8</td><td>27.8 | 55.5</td></tr>
<tr><td>PS-JP2</td><td>7.2 | 4.6</td><td>7.1 | 4.5</td><td>7.2 | 5.0</td><td>7.3 | 5.2</td><td>16.2 | 24.3</td><td>16.1 | 24.5</td><td>15.9 | 24.3</td><td>16.4 | 24.5</td><td>27.3 | 54.5</td><td>27.0 | 54.9</td><td>27.6 | 55.1</td><td>27.9 | 57.4</td></tr>
<tr><td rowspan="4">UBO-Morpher</td><td>NPP</td><td>6.6 | 4.2</td><td>6.6 | 4.2</td><td>6.6 | 4.4</td><td>7.0 | 4.5</td><td>17.2 | 27.4</td><td>17.2 | 26.6</td><td>17.4 | 26.9</td><td>17.1 | 26.8</td><td>25.5 | 51.8</td><td>25.7 | 51.4</td><td>25.8 | 51.6</td><td>26.1 | 54.3</td></tr>
<tr><td>Resized</td><td>6.6 | 4.2</td><td>6.7 | 4.1</td><td>6.6 | 4.4</td><td>6.8 | 4.4</td><td>17.2 | 26.3</td><td>17.3 | 26.2</td><td>17.2 | 25.9</td><td>17.1 | 26.3</td><td>25.2 | 49.5</td><td>24.7 | 49.0</td><td>24.9 | 49.5</td><td>25.8 | 51.2</td></tr>
<tr><td>JP2</td><td>7.0 | 5.0</td><td>7.0 | 4.9</td><td>6.9 | 4.9</td><td>6.7 | 4.9</td><td>17.6 | 28.3</td><td>17.4 | 27.9</td><td>17.7 | 27.7</td><td>17.4 | 27.7</td><td>25.8 | 53.9</td><td>26.1 | 51.9</td><td>26.1 | 52.6</td><td>26.7 | 53.6</td></tr>
<tr><td>PS-JP2</td><td>6.7 | 4.2</td><td>6.7 | 4.2</td><td>6.5 | 4.3</td><td>6.8 | 4.6</td><td>17.0 | 27.1</td><td>17.2 | 26.8</td><td>17.0 | 26.3</td><td>17.0 | 26.6</td><td>25.6 | 51.5</td><td>25.2 | 51.9</td><td>25.6 | 51.9</td><td>26.1 | 53.6</td></tr>
<tr><td rowspan="8">OpenCV</td><td rowspan="4">FaceFusion</td><td>NPP</td><td>6.6 | 4.0</td><td>6.6 | 3.7</td><td>6.7 | 4.0</td><td>6.8 | 4.3</td><td>16.2 | 24.6</td><td>16.2 | 24.1</td><td>16.3 | 25.2</td><td>16.4 | 25.3</td><td>27.1 | 54.8</td><td>26.8 | 54.8</td><td>26.7 | 54.9</td><td>27.3 | 56.2</td></tr>
<tr><td>Resized</td><td>6.8 | 4.0</td><td>6.7 | 4.1</td><td>6.7 | 4.4</td><td>6.9 | 4.7</td><td>16.3 | 23.9</td><td>16.1 | 23.7</td><td>16.4 | 24.8</td><td>16.7 | 25.7</td><td>26.8 | 53.6</td><td>25.5 | 52.8</td><td>25.8 | 52.5</td><td>26.5 | 54.6</td></tr>
<tr><td>JP2</td><td>7.2 | 4.9</td><td>7.0 | 4.9</td><td>7.0 | 5.2</td><td>7.0 | 5.2</td><td>16.8 | 25.6</td><td>16.6 | 25.4</td><td>16.8 | 26.2</td><td>16.5 | 26.4</td><td>26.9 | 53.4</td><td>26.1 | 53.1</td><td>26.4 | 53.9</td><td>27.0 | 55.0</td></tr>
<tr><td>PS-JP2</td><td>6.7 | 4.2</td><td>6.6 | 4.2</td><td>6.7 | 4.3</td><td>7.0 | 4.5</td><td>16.3 | 24.3</td><td>15.9 | 24.2</td><td>16.3 | 24.6</td><td>16.3 | 24.8</td><td>27.2 | 54.6</td><td>26.2 | 54.7</td><td>26.1 | 55.0</td><td>26.9 | 56.2</td></tr>
<tr><td rowspan="4">UBO-Morpher</td><td>NPP</td><td>6.1 | 3.5</td><td>6.0 | 3.4</td><td>6.2 | 3.7</td><td>6.4 | 4.0</td><td>16.6 | 25.7</td><td>16.8 | 25.1</td><td>16.9 | 26.5</td><td>16.6 | 26.0</td><td>25.7 | 51.4</td><td>25.1 | 50.4</td><td>25.0 | 49.4</td><td>25.4 | 52.8</td></tr>
<tr><td>Resized</td><td>6.3 | 3.6</td><td>6.3 | 3.5</td><td>6.4 | 3.8</td><td>6.5 | 4.0</td><td>16.8 | 24.2</td><td>17.0 | 24.0</td><td>17.0 | 25.3</td><td>16.8 | 25.6</td><td>24.9 | 49.9</td><td>24.1 | 48.6</td><td>24.1 | 48.9</td><td>25.1 | 50.1</td></tr>
<tr><td>JP2</td><td>6.7 | 4.0</td><td>6.5 | 4.1</td><td>6.5 | 3.9</td><td>6.4 | 4.0</td><td>17.3 | 26.4</td><td>17.3 | 26.0</td><td>17.4 | 27.7</td><td>16.9 | 26.6</td><td>25.8 | 52.7</td><td>25.7 | 52.1</td><td>25.6 | 52.0</td><td>25.7 | 52.9</td></tr>
<tr><td>PS-JP2</td><td>6.3 | 3.7</td><td>6.2 | 3.8</td><td>6.2 | 4.0</td><td>6.5 | 4.2</td><td>17.0 | 25.5</td><td>16.6 | 25.2</td><td>17.2 | 26.1</td><td>16.6 | 26.3</td><td>25.7 | 50.2</td><td>24.7 | 49.3</td><td>24.6 | 49.4</td><td>25.3 | 52.0</td></tr>
<tr><td rowspan="16">FRGCv2</td><td rowspan="8">FaceMorpher</td><td rowspan="4">FaceFusion</td><td>NPP</td><td>2.9 | 1.2</td><td>3.2 | 1.1</td><td>2.7 | 1.0</td><td>2.8 | 1.1</td><td>7.5 | 6.3</td><td>7.5 | 6.5</td><td>7.3 | 6.0</td><td>7.4 | 5.9</td><td>14.1 | 18.4</td><td>14.5 | 19.3</td><td>14.4 | 19.3</td><td>14.9 | 21.2</td></tr>
<tr><td>Resized</td><td>2.7 | 1.1</td><td>2.9 | 1.2</td><td>2.7 | 1.2</td><td>2.8 | 1.2</td><td>7.3 | 5.7</td><td>7.4 | 5.9</td><td>7.0 | 5.9</td><td>7.0 | 5.7</td><td>14.4 | 18.1</td><td>15.3 | 18.8</td><td>14.4 | 18.3</td><td>15.0 | 18.8</td></tr>
<tr><td>JP2</td><td>1.2 | 0.1</td><td>1.2 | 0.1</td><td>1.2 | 0.1</td><td>1.2 | 0.1</td><td>8.3 | 6.8</td><td>8.4 | 6.9</td><td>8.1 | 6.8</td><td>7.5 | 7.3</td><td>10.0 | 10.3</td><td>10.2 | 10.6</td><td>10.6 | 11.0</td><td>11.0 | 12.1</td></tr>
<tr><td>PS-JP2</td><td>2.8 | 1.0</td><td>3.0 | 1.0</td><td>3.0 | 1.0</td><td>2.9 | 1.0</td><td>7.4 | 6.0</td><td>7.8 | 6.7</td><td>7.9 | 6.5</td><td>7.4 | 6.2</td><td>14.4 | 19.2</td><td>15.3 | 20.7</td><td>15.0 | 20.2</td><td>15.8 | 21.1</td></tr>
<tr><td rowspan="4">UBO-Morpher</td><td>NPP</td><td>3.0 | 1.2</td><td>3.2 | 1.1</td><td>2.8 | 1.0</td><td>3.0 | 1.1</td><td>8.2 | 7.2</td><td>8.2 | 7.4</td><td>8.2 | 7.0</td><td>8.3 | 7.5</td><td>13.5 | 16.5</td><td>14.0 | 16.4</td><td>13.8 | 16.8</td><td>14.5 | 18.1</td></tr>
<tr><td>Resized</td><td>2.5 | 1.1</td><td>2.8 | 1.2</td><td>2.8 | 1.2</td><td>2.8 | 1.2</td><td>8.3 | 7.3</td><td>8.1 | 7.3</td><td>7.7 | 6.8</td><td>8.1 | 6.8</td><td>13.6 | 16.0</td><td>13.6 | 17.3</td><td>13.6 | 16.7</td><td>14.6 | 18.8</td></tr>
<tr><td>JP2</td><td>3.2 | 1.2</td><td>3.3 | 1.2</td><td>3.7 | 1.2</td><td>3.4 | 1.4</td><td>9.7 | 8.2</td><td>9.4 | 8.6</td><td>8.8 | 7.9</td><td>8.9 | 8.1</td><td>15.3 | 19.6</td><td>15.1 | 20.6</td><td>14.5 | 20.2</td><td>15.3 | 19.7</td></tr>
<tr><td>PS-JP2</td><td>2.8 | 1.0</td><td>2.8 | 1.0</td><td>2.7 | 1.0</td><td>3.0 | 1.0</td><td>8.3 | 7.4</td><td>7.9 | 7.4</td><td>8.1 | 7.3</td><td>8.1 | 7.0</td><td>14.1 | 18.2</td><td>14.1 | 19.3</td><td>14.5 | 18.3</td><td>14.8 | 19.6</td></tr>
<tr><td rowspan="8">OpenCV</td><td rowspan="4">FaceFusion</td><td>NPP</td><td>2.7 | 0.7</td><td>2.9 | 0.9</td><td>2.8 | 1.0</td><td>2.8 | 1.0</td><td>7.3 | 5.3</td><td>7.3 | 5.3</td><td>7.5 | 5.4</td><td>7.2 | 5.5</td><td>14.1 | 18.2</td><td>13.6 | 17.8</td><td>13.8 | 18.2</td><td>15.0 | 20.3</td></tr>
<tr><td>Resized</td><td>2.7 | 0.9</td><td>2.7 | 0.9</td><td>2.5 | 1.0</td><td>2.0 | 1.0</td><td>7.3 | 5.3</td><td>7.2 | 5.3</td><td>7.2 | 5.4</td><td>6.8 | 5.5</td><td>14.0 | 19.2</td><td>14.9 | 17.8</td><td>15.7 | 18.4</td><td>15.1 | 19.7</td></tr>
<tr><td>JP2</td><td>1.4 | 0.1</td><td>1.4 | 0.1</td><td>1.4 | 0.1</td><td>1.4 | 0.1</td><td>7.5 | 6.8</td><td>7.8 | 6.4</td><td>7.7 | 6.7</td><td>8.9 | 7.3</td><td>9.6 | 9.2</td><td>9.8 | 9.8</td><td>10.0 | 9.8</td><td>10.5 | 11.3</td></tr>
<tr><td>PS-JP2</td><td>2.9 | 0.9</td><td>2.9 | 0.9</td><td>2.5 | 0.7</td><td>2.4 | 1.0</td><td>7.2 | 5.3</td><td>7.0 | 5.7</td><td>7.2 | 5.5</td><td>7.3 | 5.3</td><td>14.9 | 20.0</td><td>14.0 | 18.6</td><td>14.1 | 19.8</td><td>15.5 | 21.3</td></tr>
<tr><td rowspan="4">UBO-Morpher</td><td>NPP</td><td>2.7 | 0.7</td><td>2.8 | 0.7</td><td>2.7 | 0.7</td><td>2.8 | 0.9</td><td>7.5 | 6.3</td><td>7.5 | 6.0</td><td>7.7 | 6.5</td><td>7.9 | 6.5</td><td>13.1 | 16.5</td><td>12.7 | 15.7</td><td>13.5 | 16.3</td><td>14.3 | 17.7</td></tr>
<tr><td>Resized</td><td>2.7 | 0.9</td><td>2.8 | 0.9</td><td>2.3 | 0.9</td><td>2.7 | 0.9</td><td>7.5 | 6.7</td><td>7.9 | 6.4</td><td>7.9 | 6.7</td><td>7.9 | 6.0</td><td>13.1 | 16.9</td><td>13.6 | 16.9</td><td>13.4 | 16.5</td><td>14.8 | 18.3</td></tr>
<tr><td>JP2</td><td>3.0 | 1.2</td><td>2.9 | 1.0</td><td>2.9 | 1.0</td><td>3.2 | 1.1</td><td>8.6 | 7.4</td><td>8.4 | 7.2</td><td>8.4 | 7.7</td><td>9.1 | 8.7</td><td>14.1 | 19.6</td><td>14.3 | 18.2</td><td>14.4 | 19.3</td><td>14.3 | 21.3</td></tr>
<tr><td>PS-JP2</td><td>2.5 | 0.9</td><td>2.7 | 0.9</td><td>2.5 | 0.7</td><td>2.7 | 0.9</td><td>7.7 | 6.4</td><td>7.9 | 6.7</td><td>8.1 | 6.4</td><td>7.5 | 6.3</td><td>13.5 | 19.4</td><td>13.0 | 17.7</td><td>13.6 | 18.1</td><td>14.8 | 18.4</td></tr>
</table>

for the higher quality morphs (FaceFusion and UBO-Morpher) and even comparably good for the lower quality morphs. However, on the FERET database, its performance even falls behind that of the Eyedea-based MAD algorithm.

Since the use of grayscale probe images is not expected to significantly impact the recognition performance of face recognition systems based on deep face representations it is valid to assume that the proposed MAD approaches achieve equal detection performance in case colour probe images are processed. The same holds for MAD algorithms which explicitly perform a grayscale conversion prior to the feature extraction, e.g., MAD based on LBP or BSIF feature extractors.

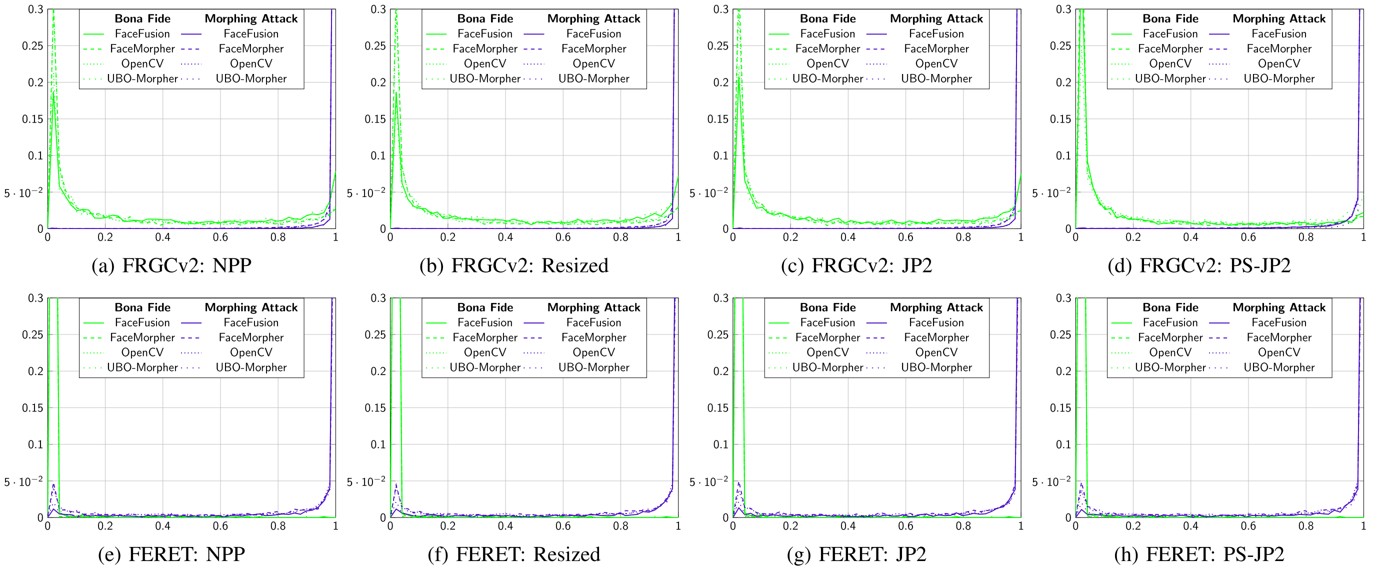
A general pattern observable in the results is that morphs with higher quality are more difficult to detect. FaceFusion and the UBO-Morpher both include automatic post-processing, which replaces the artefact-rich region outside the face and therefore generate a higher error rate during MAD.[9] Although the best results are achieved if for training and test sets the same morphing tool is used (which indicates slight overfitting to the characteristics of the morphing process), the influence of the morphing tool used for training is quite limited.

B. Robustness Against Image Post-Processing

In this subsection, the robustness of MAD algorithms based on deep face representations with respect to post-processing of the reference images is investigated. All four post-processings presented in Section III-E are separately used for training and testing. Since we have already seen that the morphing tool used for training is of minor importance and that higher quality morphs (created with FaceFusion and UBO-Morpher) are more difficult to detect, we use FaceMorpher and OpenCV for training, and FaceFusion and UBO-Morpher for testing. The other MAD algorithms considered in the previous section will not be further considered, because they cannot achieve comparable performance.

The corresponding results are shown in Table VI. As in the previous experiment, the MAD algorithm based on deep face represnetations extracted using ArcFace is far ahead in detection performance, followed by the MAD algorithm based on Eyedea. Training on post-processed images has no noticeable influence on the detection performance of the algorithms. Testing on post-processed images can have a slight effect on detection performance. In general, the proposed

[9] In addition, the outer region is replaced with that from the attacker's image which is expected to increase the similarity in comparison with a probe image from the attacker.



(a) FRGCv2: NPP (b) FRGCv2: Resized (c) FRGCv2: JP2 (d) FRGCv2: PS-JP2

(e) FERET: NPP (f) FERET: Resized (g) FERET: JP2 (h) FERET: PS-JP2

Fig. 10. PDFs of the decision score of an MAD based on ArcFace features for different databases and post-processings, when trained on different morphing attack algorithms. Low scores (close to 0) indicate to bona fide authentication attempts and high scores (close to 1) indicate MAs.

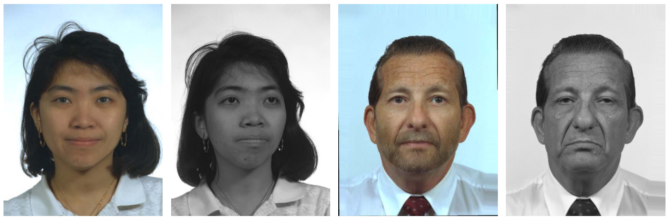
MAD algorithms are not severely influenced by the tested post-processings. Presumably, this is due to the extraction of the deep face representations by the neural networks which were trained for a high independence from image properties.

C. Analysis of Score Distributions

As a final investigation, score distributions of the MAD algorithm based on the ArcFace feature extraction are analyzed in more detail. Since this MAD algorithm achieves by far the best performance, the other two deep face representation extractors (Eyedea and Facenet) are not considered here. Since post-processing has no influence during training, only algorithms trained on NPP images are considered. Trained SVMs generate a normalized MAD score in the range $[0, 1]$ where low scores, i.e., scores close to 0, indicate bona fide authentication attempts and high scores, i.e., scores close to 1, refer to MAs. Figure 10 shows the resulting probability density functions. For each subfigure, training was performed with each morphing tool separately (the tool name is given in the legend of the plot), while testing was performed considering MAs using all morphing algorithms together. The database and post-processing used for the evaluation are given in the caption of the respective subfigure.

The observations correspond to the findings of the previous experiments. Within a database, the score distributions are very similar, regardless of which morphing tool is used for training and which post-processing is used for the test set. This further suggests that the use of a variety of different morphing tools in the training stage is not expected to yield significant improvement in terms of detection performance. Generally, scores which were generated by the evaluation on FERET can be separated better than those stemming from FRGCv2.

It is noticeable that when evaluated on FRGCv2, some bona fide samples are very clearly, i.e., with rather high scores, misclassified as MAs, visible in Figure 10 (a) to (d) as a peak for bona fides at 1. On the contrary, on FERET, misclassifications occur mostly for MAs, which can be seen in Figure 10 (e) to (h) as small bumps in the MA score distributions close to 0. Examples for the above mentioned errors are given in Figure 11 and Figure 12. If the morph is of good quality and very similar to the sample, the MAD algorithm can no longer detect the MA. For example, in the samples shown in Figure 11 it is difficult even for an experienced observer to detect the MA. Examples for wrongly classified bona fide authentication attempts are given in Figure 12. The cause of this error is obvious. In these examples, the strong variance in facial expression and the presence of headgear in the probe images are particularly noticeable. However, such a variance can also be expected in a realistic border control scenario.



(a) Reference 1 (b) Probe 1 (c) Reference 2 (d) Probe 2

Fig. 11. Attack presentation classification error examples on FERET.

Due to the high certainty with which the algorithm assigns some samples to the wrong category, it is hardly possible to correct these errors by adjusting the MAD decision threshold value. Still, for both databases the score distributions of MAs and bona fide authentication attempts are very clearly separated. Thus, the error rates are quite stable within a considerable range of MAD decision threshold values (operating points), which makes the MAD algorithm very suitable for a practical application.

D. NIST Face Recognition Vendor Test MORPH

The NIST FRVT MORPH test provides ongoing independent testing of prototype face MAD methods. The evaluation is

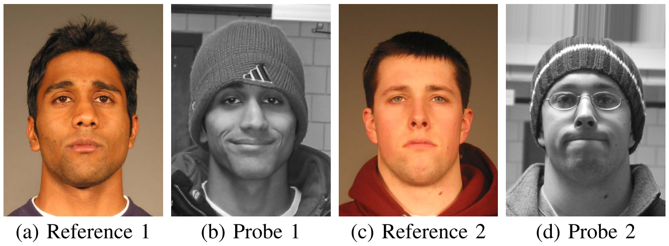


Fig. 12. Bona fide presentation classification error examples on FRGCv2.

designed to obtain commonly measured assessment of morph detection capability to inform developers and end-users. The test opened in June 2018, and, at the time of this writing, NIST has since received twelve different MAD algorithms. The latest report [22] includes morph detection performance results over thirteen datasets. On various unseen datasets where differential MAD was evaluated the proposed method achieved performance rates comparable to those presented in this work. Further, it outperformed all other methods by orders of magnitude, including implementations of single image MAD methods based on texture descriptors, forensic image analysis, and deep learning, i.e., [7], [18]–[20], [23], [38], [39], [56], as well as differential MAD methods employing texture descriptors and facial landmarks, i.e., [7], [11]. For detailed results the interested reader is referred to [22]. The submitted version of the proposed MAD method uses ArcFace for the extraction of deep face representations. For the training of the corresponding SVM, we utilized bona fide and morphed reference images from the FERET and FRGCv2 databases. In order to achieve high robustness morphed reference images generated by all morphing tools used in this work were considered in the training stage.

VI. Conclusion

Based on the experiments conducted in this work, the following conclusions are reached:

- *Detection performance*: the detection performance achieved by MAD based on deep face representations is promising and highly robust with respect to image post-processing, i.e., image compression, image resizing and even print-scan transformation. This is a clear advantage over MAD based on texture descriptors, which is typically quite sensitive to post-processing, particularly in more challenging scenarios. Moreover, the detection performance does not significantly depend on the post-processing applied to the training set, so that no scanned images are necessary for training.
- *Heterogeneous morphing algorithms:* morphs generated by morphing algorithms which produce obvious artefacts, e.g., clearly visible ghost artefacts, are generally detected with higher accuracy. Furthermore, the recognition performance slightly degrades if training and evaluation sets contain morphs generated by different morphing algorithms.
- *Heterogeneous databases:* if training and testing is conducted on heterogeneous face image databases which contain face images with different conditions, e.g., variations in pose and lightning, detection performance is negatively affected. On databases obtained from subsets of the publicly available FERET and the FRGCv2 face database, experiments revealed higher detection accuracy on the FERET subset in which probe images only contain slight variations in expression and pose as opposed to the FRGCv2 subset, which additionally comprises probe images with variations in lightning and focus. It can be concluded that strong variations in lightning and focus of probe images represent especially challenging conditions for differential MAD.
- *Machine learning-based classifiers:* among the tested machine learning-based classifiers, i.e., AdaBoost, Gradient Boosting, Random Forest and Support Vector Machine (SVM), SVM-based classifiers generally revealed most competitive detection performance across the vast majority of conducted experiments.
- *Commercial vs. open-source:* while commercial face recognition algorithms frequently outperform corresponding open-source implementations, this is not necessarily the case for MAD. Precisely, for the task of MAD, deep face representations obtained from open-source algorithms, e.g., ArcFace, might be better suited, compared to deep features extracted by commercial face recognition systems.

Furthermore, this work underlines the need for realistic databases. Not only the quality of the reference, but also the quality of the probes has a strong influence on the detection performance of the MAD algorithms. Therefore, for the development of algorithms, which are deployable in a real world scenario, it is necessary to test on realistic data. However, there is still the problem that the exchange of databases is difficult due to privacy regulations.

Acknowledgment

The UBO-Morpher as well as the Demorphing implementation was kindly provided by the University of Bologna.

References

[1] U. Scherhag, C. Rathgeb, J. Merkle, R. Breithaupt, and C. Busch, “Face recognition systems under morphing attacks: A survey,” *IEEE Access*, vol. 7, pp. 23012–23026, 2019.

[2] M. Ferrara, A. Franco, and D. Maltoni, “On the effects of image alterations on face recognition accuracy,” in *Face Recognition Across the Imaging Spectrum*. Cham, Switzerland: Springer, 2016, pp. 195–222.

[3] D. J. Robertson, A. Mungall, D. G. Watson, K. A. Wade, S. J. Nightingale, and S. Butler, “Detecting morphed passport photos: A training and individual differences approach,” *Cognit. Res. Princ. Implications*, vol. 3, no. 1, pp. 1–11, Dec. 2018.

[4] M. Ferrara, A. Franco, and D. Maltoni, “The magic passport,” in *Proc. IEEE Int. Joint Conf. Biometrics*, Sep. 2014, pp. 1–7.

[5] M. Gomez-Barrero, C. Rathgeb, U. Scherhag, and C. Busch, “Is your biometric system robust to morphing attacks?” in *Proc. 5th Int. Workshop Biometrics Forensics (IWBF)*, Apr. 2017, pp. 1–6.

[6] M. Gomez-Barrero, C. Rathgeb, U. Scherhag, and C. Busch, “Predicting the vulnerability of biometric systems to attacks based on morphed biometric information,” *IET Biometrics*, vol. 7, no. 4, pp. 333–341, Jul. 2018.

[7] U. Scherhag, C. Rathgeb, and C. Busch, “Towards detection of morphed face images in electronic travel documents,” in *Proc. 13th IAPR Workshop Document Anal. Syst. (DAS)*, 2018, pp. 187–192.

[8] P. Phillips *et al.*, “Overview of the face recognition grand challenge,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 947–954.

[9] U. Scherhag, D. Budhrani, M. Gomez-Barrero, and C. Busch, “Detecting morphed face images using facial landmarks,” in *Proc. Int. Conf. Image Signal Process. (ICISP)*. Cham, Switzerland: Springer, 2018, pp. 444–452.

[10] A. Martinez and R. Benavente, “The AR face database,” Comput. Vis. Center, Barcelona, Spain, Tech. Rep. 24, Jun. 1998.

[11] N. Damer *et al.*, “Detecting face morphing attacks by analyzing the directed distances of facial landmarks shifts,” in *Proc. 40th German Conf. Pattern Recognit. (GCPR)*, 2018, pp. 518–534.

[12] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, “The FERET database and evaluation procedure for face-recognition algorithms,” *Image Vis. Comput.*, vol. 16, no. 5, pp. 295–306, Apr. 1998.

[13] M. Ferrara, A. Franco, and D. Maltoni, “Face demorphing,” *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 4, pp. 1008–1017, Apr. 2018.

[14] M. Ferrara, A. Franco, and D. Maltoni, “Face demorphing in the presence of facial appearance variations,” in *Proc. 26th Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2018, pp. 2365–2369.

[15] W. Gao, B. Cao, S. Shan, D. Zhou, X. Zhang, and D. Zhao, “The CAS-PEAL large-scale chinese face database and baseline evaluations,” Chin. Acad. Sci., Beijing, China, Tech. Rep. JDL-TR-04-FR-001, May 2004.

[16] F. Peng, L.-B. Zhang, and M. Long, “FD-GAN: Face de-morphing generative adversarial network for restoring accomplice’s facial image,” *IEEE Access*, vol. 7, pp. 75122–75131, 2019.

[17] A. Makrushin, T. Neubert, and J. Dittmann, “Automatic generation and detection of visually faultless facial morphs,” in *Proc. 12th Int. Joint Conf. Comput. Vis., Imag. Comput. Graph. Theory Appl. (SCITEPRESS)*, 2017, pp. 39–50.

[18] M. Ferrara, A. Franco, and D. Maltoni, “Face morphing detection in the presence of printing/scanning and heterogeneous image sources,” *CoRR*, vol. abs/1901.08811, pp. 1–23, Jan. 2019. [Online]. Available: http://arxiv.org/abs/1901.08811

[19] R. Raghavendra, K. B. Raja, S. Venkatesh, and C. Busch, “Transferable deep-CNN features for detecting digital and print-scanned morphed face images,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1822–1830.

[20] R. Ramachandra, S. Venkatesh, K. Raja, and C. Busch, “Towards making morphing attack detection robust using hybrid scale-space colour texture features,” in *Proc. IEEE 5th Int. Conf. Identity, Secur., Behav. Anal. (ISBA)*, Jan. 2019, pp. 1–8.

[21] U. Scherhag, R. Raghavendra, K. B. Raja, M. Gomez-Barrero, C. Rathgeb, and C. Busch, “On the vulnerability of face recognition systems towards morphed face attacks,” in *Proc. 5th Int. Workshop Biometrics Forensics (IWBF)*, Apr. 2017, pp. 1–6.

[22] M. Ngan, P. Grother, K. Hanaoka, and J. Kuo, “Face recognition vendor test (FRVT) part 4: MORPH performance of automated face morph detection,” Nat. Inst. Technol., Gaithersburg, MD, USA, Tech. Rep. NISTIR 8292, 2020.

[23] R. Raghavendra, K. B. Raja, and C. Busch, “Detecting morphed face images,” in *Proc. IEEE 8th Int. Conf. Biometrics Theory, Appl. Syst. (BTAS)*, Sep. 2016, pp. 1–7.

[24] C. Kraetzer, A. Makrushin, T. Neubert, M. Hildebrandt, and J. Dittmann, “Modeling attacks on photo-ID documents and applying media forensics for the detection of facial morphing,” in *Proc. 5th ACM Workshop Inf. Hiding Multimedia Secur. (IHMMSec)*, 2017, pp. 21–32.

[25] C. Seibold, W. Samek, A. Hilsmann, and P. Eisert, “Detection of face morphing attacks by deep learning,” in *Digital Forensics and Watermarking*. Cham, Switzerland: Springer, 2017, pp. 107–120.

[26] T. Ojala, M. Pietikäinen, and D. Harwood, “A comparative study of texture measures with classification based on featured distributions,” *Pattern Recognit.*, vol. 29, no. 1, pp. 51–59, Jan. 1996.

[27] J. Kannala and E. Rahtu, “BSIF: Binarized statistical image features,” in *Proc. 21st Int. Conf. Pattern Recognit. (ICPR)*, Nov. 2012, pp. 1363–1366.

[28] U. Scherhag *et al.*, “Biometric systems under morphing attacks: Assessment of morphing techniques and vulnerability reporting,” in *Proc. Int. Conf. Biometrics Special Interest Group (BIOSIG)*, Sep. 2017, pp. 1–7.

[29] U. Scherhag, C. Rathgeb, and C. Busch, “Performance variation of morphed face image detection algorithms across different datasets,” in *Proc. Int. Workshop Biometrics Forensics (IWBF)*, Jun. 2018, pp. 1–6.

[30] U. Scherhag, C. Rathgeb, and C. Busch, “Morph deterction from single face image: A multi-algorithm fusion approach,” in *Proc. 2nd Int. Conf. Biometric Eng. Appl. (ICBEA)*, 2018, pp. 6–12.

[31] L. Spreeuwers, M. Schils, and R. Veldhuis, “Towards robust evaluation of face morphing detection,” in *Proc. 26th Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2018, pp. 1027–1031.

[32] N. Damer, A. M. Saladie, A. Braun, and A. Kuijper, “MorGAN: Recognition vulnerability and attack detectability of face morphing attacks created by generative adversarial network,” in *Proc. IEEE 9th Int. Conf. Biometrics Theory, Appl. Syst. (BTAS)*, Oct. 2018, pp. 1–10.

[33] R. Raghavendra, K. Raja, S. Venkatesh, and C. Busch, “Face morphing versus face averaging: Vulnerability and detection,” in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Oct. 2017, pp. 555–563.

[34] L. Wandzik, G. Kaeding, and R. V. Garcia, “Morphing detection using a general-purpose face recognition system,” in *Proc. 26th Eur. Signal Process. Conf. (EUSIPCO)*, 2018, pp. 1012–1016.

[35] A. Asaad and S. Jassim, “Topological data analysis for image tampering detection,” in *Proc. 10th Int. Workshop Digit. Forensics Watermarking (IWDW)*. Cham, Switzerland: Springer, 2017, pp. 136–146.

[36] S. Jassim and A. Asaad, “Automatic detection of image morphing by topology-based analysis,” in *Proc. 26th Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2018.

[37] A. Agarwal, R. Singh, M. Vatsa, and A. Noore, “SWAPPED! Digital face presentation attack detection via weighted local magnitude pattern,” in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Oct. 2017, pp. 659–665.

[38] L. Debiasi, U. Scherhag, C. Rathgeb, A. Uhl, and C. Busch, “PRNU-based detection of morphed face images,” in *Proc. Int. Workshop Biometrics Forensics (IWBF)*, Jun. 2018, pp. 1–7.

[39] L. Debiasi, C. Rathgeb, U. Scherhag, A. Uhl, and C. Busch, “PRNU variance analysis for morphed face image detection,” in *Proc. IEEE 9th Int. Conf. Biometrics Theory, Appl. Syst. (BTAS)*, Oct. 2018, pp. 1–9.

[40] U. Scherhag, L. Debiasi, C. Rathgeb, C. Busch, and A. Uhl, “Detection of face morphing attacks based on PRNU analysis,” *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 1, no. 4, pp. 302–317, Oct. 2019.

[41] L. Debiasi *et al.*, “On the detection of GAN-based face morphs using established morph detectors,” in *Proc. 20th Int. Conf. Image Anal. Process. (ICIAP)*, 2019, pp. 345–356.

[42] L.-B. Zhang, F. Peng, and M. Long, “Face morphing detection using Fourier spectrum of sensor pattern noise,” in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2018, pp. 1–6.

[43] M. Hildebrandt, T. Neubert, A. Makrushin, and J. Dittmann, “Benchmarking face morphing forgery detection: Application of stirtrace for impact simulation of different processing steps,” in *Proc. 5th Int. Workshop Biometrics Forensics (IWBF)*, Apr. 2017, pp. 1–6.

[44] C. Seibold, A. Hilsmann, and P. Eisert, “Reflection analysis for face morphing attack detection,” in *Proc. 26th Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2018, pp. 1–6.

[45] C. Seibold, W. Samek, A. Hilsmann, and P. Eisert, “Accurate and robust neural networks for security related applications exampled by face morphing attacks,” in *Proc. Comput. Vis. Pattern Recognit.*, 2018, pp. 1–16.

[46] *Machine Readable Travel Documents—Part 9: Deployment of Biometric Identification and Electronic Storage of Data in MRTDs (7th Edition)*, document ICAO 9303, ICAO, International Civil Aviation Organization, 2015.

[47] D. E. King, “Dlib-ml: A machine learning toolkit,” *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, Dec. 2009.

[48] A. W. Yip and P. Sinha, “Contribution of color to face recognition,” *Perception*, vol. 31, no. 8, pp. 995–1003, Aug. 2002.

[49] R. Ranjan *et al.*, “Deep learning for understanding faces: Machines may be just as good, or better, than humans,” *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 66–83, Jan. 2018.

[50] *EU-eMRTD Specification*, European Commission, Brussels, Belgium, 2018.

[51] *Best Practice Technical Guidelines for Automated Border Control ABC Systems*, FRONTEX, Warsaw, Poland, Sep. 2015.

[52] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “ArcFace: Additive angular margin loss for deep face recognition,” 2018, *arXiv:1801.07698*. [Online]. Available: http://arxiv.org/abs/1801.07698

[53] F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: A unified embedding for face recognition and clustering,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.

[54] *ISO/IEC JTC1 SC37 Biometrics, Information Technology—Biometric Presentation Attack Detection—Part 3: Testing and Reporting*, document ISO ISO/IEC IS 30107-3:2017, International Organization for Standardization, Geneva, Switzerland, 2017.

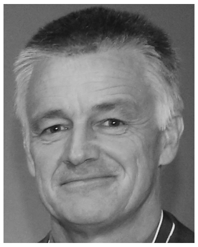
[55] Z.-H. Feng, J. Kittler, M. Awais, P. Huber, and X.-J. Wu, “Wing loss for robust facial landmark localisation with convolutional neural networks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2235–2245.

[56] U. Scherhag, L. Debiasi, C. Rathgeb, C. Busch, and A. Uhl, “Detection of face morphing attacks based on PRNU analysis,” in *Proc. Trans. Biometrics, Behav., Identity Sci. (T-BIOM)*, 2019, pp. 302–317.

Ulrich Scherhag received the B.Eng. degree in electrical engineering from Duale Hochschule Baden-Württemberg, Mannheim, in 2012, and the M.Sc. degree in computer science, IT-security from Hochschule Darmstadt in 2016. Since 2016, he has been a Ph.D. Student Member with the da/sec, Center for Research in Security and Privacy (CRISP). His current research interests include presentation attack detection and morphed face detection. He is a member of the European Association for Biometrics (EAB). He was a recipient of the CAST Award IT-Security, for his M.Sc. degree, in 2016. He is a Reviewer for the International Conference of the Biometrics Special Interest Group (BIOSIG) and IEEE ACCESS.

Christian Rathgeb is currently a Senior Researcher with the Faculty of Computer Science, Hochschule Darmstadt (HDA), Germany. He is a Principal Investigator with the Center for Research in Security and Privacy (CRISP). He has coauthored over 100 technical articles in the field of biometrics. His research interests include pattern recognition, iris and face recognition, security aspects of biometric systems, secure process design, and privacy enhancing technologies for biometric systems. He is a member of the European Association for Biometrics (EAB) and an Editorial Board Member of IET Biometrics (IET BMT). He was a recipient of the EAB—European Biometrics Research Award 2012, the Austrian Award of Excellence 2012, Best Poster Paper Awards (IJCB'11, IJCB'14, and ICB'15), and the Best Paper Award Bronze (ICB'18). He is the Program Chair of the International Conference of the Biometrics Special Interest Group (BIOSIG). He has served for various program committees and conferences (e.g., ICB, IJCB, BIOSIG, and IWBF) and journals (e.g., the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY (TIFS), the IEEE TRANSACTIONS ON BIOMETRICS, BEHAVIOR, AND IDENTITY SCIENCE (TBIOM), and IET BMT) as a reviewer.

Johannes Merkle graduated in mathematics from the University in Frankfurt in 1995. After working there for three years in the field of cryptography, he received the Ph.D. degree in mathematics in 2000. Since 1998, he has been working as a Security Consultant with secunet Security Networks, where he focussed on cryptography, public key infrastructures, and biometric cryptosystems. He has published several research articles on biometric template protection, fingerprint recognition, and elliptic curve cryptography.



Christoph Busch is currently a member of the Department of Information Security and Communication Technology (IIK), Norwegian University of Science and Technology (NTNU), Norway. He holds a joint appointment with the Computer Science Faculty, Hochschule Darmstadt (HDA), Germany. Further, he has been Lecturing a course on biometric systems with the Technical University of Denmark (DTU) since 2007. He has coauthored more than 400 technical articles and has been a speaker at international conferences. He served for various program committees (NIST IBPC, ICB, ICHB, BSI-Congress, GI-Congress, DACH, WEDELMUSIC, and EUROGRAPHICS) and served for several conferences, journals, and magazines, as a reviewer (e.g., ACM-SIGGRAPH, ACM-TISSEC, the IEEE COMPUTER GRAPHICS AND APPLICATIONS (CG&A), the IEEE TRANSACTIONS ON SIGNAL PROCESSING, the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, the IEEE TRANSACTIONS ON PATTERN ANALYSIS, *Machine Intelligence* and the Elsevier Journal *Computers and Security*). He is also an Appointed Member of the Editorial Board of the *IET Journal on Biometrics* and the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY (TIFS) journal. Furthermore, he chairs the Biometrics Working Group, TeleTrusT Association, and the German Standardization Body on Biometrics (DIN-NIA37). He is a Convenor of WG3 in ISO/IEC JTC1 SC37 on Biometrics and an Active Member of CEN TC 224 WG18.