Deep Face Representations for Differential Morphing Attack Detection

Ulrich Scherhag, Christian Rathgeb, Johannes Merkle, Christoph Busch

The following paper is a pre-print. The publication is currently under review for IEEE Transactions on Information Forensics and Security (TIFS).

***Abstract*—The vulnerability of facial recognition systems to face morphing attacks is well known. Many different approaches for morphing attack detection have been proposed in the scientific literature. However, the morphing attack detection algorithms proposed so far have only 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 Face Recognition Vendor Test MORPH show that the submitted MAD algorithms lack robustness and performance when considering unseen and challenging datasets.**

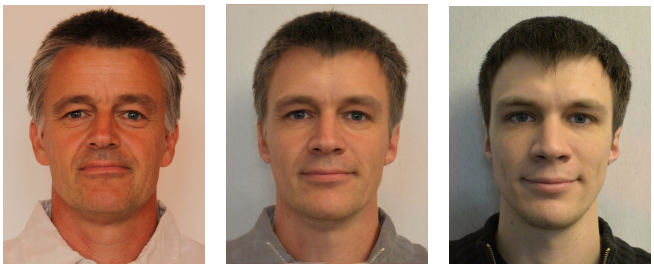
In this work, subsets of the FERET and FRGCv2 face databases are used to create a large realistic database for training and testing of morphing attack detection algorithms, containing a large number of ICAO-compliant bona fide facial images, corresponding unconstrained probe images, and morphed images created with four different 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 morphing attack detection 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. Realistic looking morph images can be generated by unskilled persons using readily available tools [1]. An example of a morphed facial image is shown as part of Figure 1.

U. Scherhag is with the da/sec - Biometrics and Internet Security Research Group, Hochschule Darmstadt, Germany
C. Rathgeb is with the da/sec - Biometrics and Internet Security Research Group, Hochschule Darmstadt, Germany, and the secunet Security Networks AG, Essen, Germany
J. Merkle is with the secunet Security Networks AG, Essen, Germany
C. Busch is with the da/sec - Biometrics and Internet Security Research Group, Hochschule Darmstadt, Germany



(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, and the accomplice 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 potential to launch such a Morphing Attack (MA) in practice was showcased in Germany 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].

Many approaches for Morphing Attack Detection (MAD) algorithms have already been proposed in the literature. Most of these are *single image 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 (with the notable exception of [15], [16], [17] and [18]) do not consider the post-processings of the images (e.g. print-scan transformation, strong compression) which can occur in real-world scenarios and may drastically reduce detectable artefacts from the morphing process. In addition, all previous publications on differential MAD use probe images that don't exhibit a realistic intra-subject vari-

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

TABLE I
OVERVIEW OF MOST RELEVANT DIFFERENTIAL MAD ALGORITHMS.

Ref.	Approach	Category	Morphing method	Source face database	Post-processing	Remarks
[4]	Differential BSIF + SVM	Feature comparison	triangulation + blending	FRGCv2 [5]	-	-
[6]	Landmarks	Feature comparison	triangulation + blending	ARface [7]	-	-
[8]	Directed Distances of Landmarks	Feature comparison	triangulation + blending (+ swapping)	FERET [9]	-	-
[10]	Demorphing	Morphing reversion	GIMP/GAP	ARface [7]	-	-
[11]	Demorphing	Morphing reversion	GIMP/GAP	ARface [7], CAS-PEAL-R1 [12]	-	CAS-PEAL-R1 contains images with pose variations
[13]	DNN-Demorphing	Morphing reversion	triangulation + blending [14]	in-house	-	-

ance, e.g. in pose, facial 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 morph images generated with the same tools as the images in the training set. As a consequence, the low error rates reported in these publications may simply reflect an overfitting to the specific and unrealistic characteristics of the images used. This suspicion is supported by the recent results of the NIST Face Recognition Vendor Test MORPH [19] test, where none of the submitted algorithms achieved satisfactory performance for all data sets.

This paper is structured as follows. Previous results 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, a conclusion is given 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

The single-image MAD approaches can be categorized into three classes: *Texture descriptors*, e.g., in [20], *forensic image analysis*, e.g., in [21], and *methods based on deep neural networks*, e.g., in [22]. These differ in the artefacts they can potentially detect.

Texture descriptors, e.g. Local Binary Patterns (LBP) [23] or Binarized Statistical Image Features (BSIF) [24] attempt to extract discriminative information from image 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 appear more frequently, examples of those artifacts are shown in [25]. 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 different ways [18], [20], [4], [26], [27], [28], [17], [29], [30], [31], [32].

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 of the images changes. It was shown in [33], [34], [35], [36], [37] that these changes can be used for detection. Under the assumption that the images are intermediately stored during the morph creation process with lossy compression algorithms, double compression artefacts can be analyzed [14], [38]. Furthermore, inconsistencies in the image, for example inconsistent illumination [39] or color values, can be evaluated.

Deep neural networks can be used to detect morphs in two different ways. Firstly, a new neural network is trained or an existing neural network is re-trained [16], [22], [40] for the task of morphing detection. 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 specific artefacts. The second method is to use the feature vectors (embeddings) of existing deep nets [29]. Since the neural network was not 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 summarized in Table I. Differential MAD can be divided into two categories. The first category are algorithms that compare two feature vectors. For example, single image algorithms can be extended to differential algorithms by comparing the feature vectors of reference and probe, e.g., by estimating differences between feature vectors extracted from trusted live captures and potential morphs [4]. The additional information of the bona fide live capture might improve the performance and robustness of the detection algorithm. Further, algorithms explicitly utilizing the additional information have been introduced in [6], [8], where the distances between facial landmarks are estimated.

The second category contains algorithms that try to reverse the process of morphing. The assumption here is that if two subjects are represented in the morphed image and the

trusted live-capture image 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 [10], [11] so-called de-morphing was proposed, an approach where reversion is done explicitly. In addition to the explicit de-morphing approach, de-morphing based on a Generative Adversarial Network (GAN) is proposed in [13].

III. Creation of MAD Database

For the development and testing of MAD algorithms, bona fide and morphed reference images are needed. In order to resemble passport photos, these images should meet the requirements of the ICAO passport photo quality standards [41]. 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 required. 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 10 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.

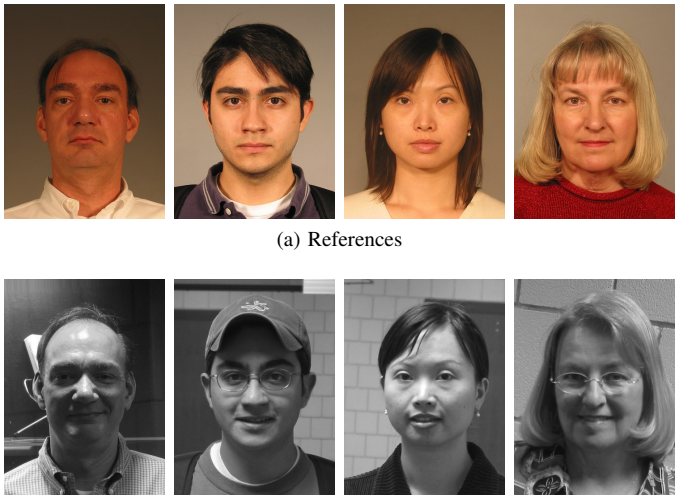
Unfortunately, there is yet no public database available that contains facial images which exhibit all of the mentioned properties. Therefore, we decided to set up a new MAD database based on existing face image databases. In this section, the steps taken for this, including the selection of the images from public face image databases, 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 built 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 passport-style images must be available (one of which used for generating morphs 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 [9] and FRGCv2 [5] 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.



(a) References

(b) Probes

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

B. Selection of Image Candicates

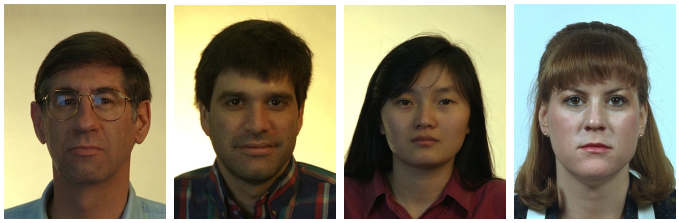
For each of these databases, we created two lists of images, representing the passport photographs and the unconstrained recordings.

The list of passport photographs contains all images complying with ICAO-requirements [41] 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.

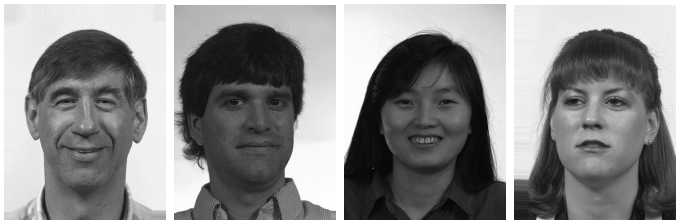
From the remaining images, those suitable as probe images to resemble the border crossing capture process were selected. The face should be recognizable, but can be only partially illuminated and slightly out of focus. The selected images of the FERET database yield a variance in facial expression and pose (slight rotations), for FRGCv2, the selected images yield a variance in facial expression, background, illumination and sharpness. In order to simulate the conditions at the eGate, the probe images are converted to grayscale. Examples of the reference and probe images of both databases are shown in Figure 2 and 3.

C. Selection of the Images

For each subject, we selected from the list of passport photographs a set of bona fide reference image and a set of input samples for generating morphs, as well as a set of probe images from the list of unconstrained recordings. Where possible, we ensured that the bona fide images were



(a) References



(b) Probes

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

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	Passport Images		Probes
				Bona Fide	Morphs	
FERET	529	329	200	529	4 · 529	791
FRGCv2	533	302	231	984	4 · 964	1,726

disjoint from the input images used for the morphing process. In addition, we required that the samples selected from the list of passport photos (bona fide or morphing input) should, if possible, be captured in a different recording session than the probe images (with maximum time difference), in order to reflect the temporal difference between passport image and live capture in border control scenarios. For FERET, we selected one bona fide image, one morphing input image and up to two probe images per subject. For FRGCv2, we chose two bona fide images, two morphing input images and up to 5 probe images per subject.

For the creation of morphs, we defined a list of image pairs to be morphed. Thereby, we ensured that only subjects of the same gender were morphed, and, to avoid obvious artefacts, that only one of the input images showed glasses. Furthermore, each sample was used only for one morph. The key figures of the resulting database are listed in Table II.

D. Morphing Process

Four different 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. Optionally, the corresponding landmarks of upper and lower lips can be reduced as described in [14] to avoid artefacts at closed mouths. The created morphs have a high quality and low to no visible artefacts. An example is 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 localisation. 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. An example is shown in Figure 4c.
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 for this algorithm landmarks are positioned at the edge of the image, which are also used to create the morphs. Thus, in contrast to FaceMorpher, the edge does not consist of an averaged image, but like the rest of the image, of morphed triangles. However, strong artefacts outside the face area can be observed, which is mainly due to missing landmarks. An example is shown in Figure 4d.
4) **UBO**, the morphing tool of University of Bologna, as used, e.g., in [10]. This algorithm receives the two input images as well as the corresponding landmarks. 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, visible edges may appear at the transitions. An example is shown in Figure 4e.

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 the whole database, the bona fide and morphed face images are normalized to meet the ICAO-requirements for passport images [41]. The resulting images are 720×960 pixels.

E. Post-Processing

Images that have been captured for use in a passport can go through various processing steps before they are embedded in the 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. In the text below referred to as *NPP* (no post-processing). This serves as baseline.
Resized: The resolution of the images is reduced by half, reflecting the average size of a passport image.

[2]www.wearemoment.com/FaceFusion

[3]github.com/alyssaq/face_morpher

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

(a) Subject 1 (b) FaceFusion (c) FaceMorpher (d) OpenCV (e) UBO (f) Subject 2

Fig. 4. Examples for a morphed face images from all four algorithms

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

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

In the text below referred to as *Resized*. 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 [42]. 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. In the text below referred to as *JP2*.

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 by half and then 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. In the text below referred to as *PS-JP2*.

F. Validation of Attack Potential

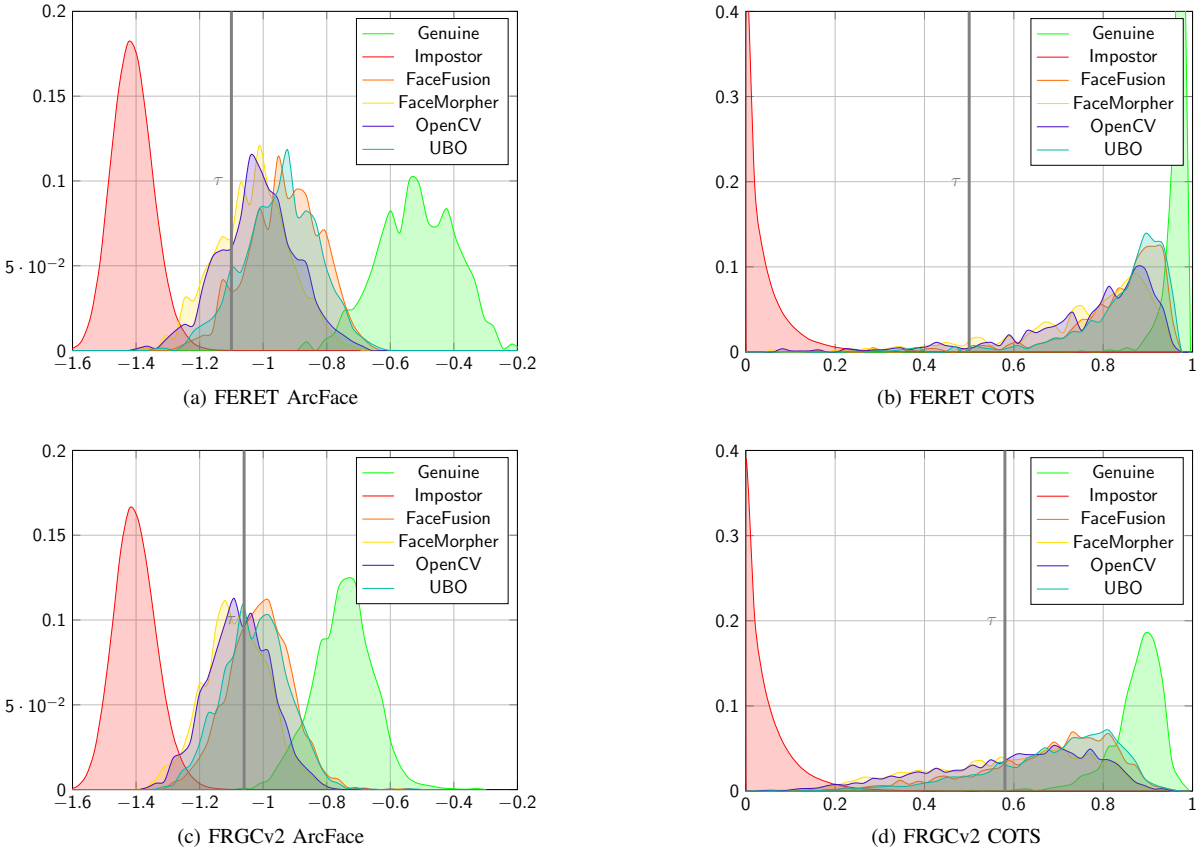
The characteristics of our MAD database differs considerably from other databases used in publications on MA and MAD. In particular, the intra-subject variation is much higher in our database due to the selection of unconstrained recordings as probe images. While this approach ensures that our database is more eligible to simulate real-world scenarios, it is perfectly valid to ask whether the unconstrained probe images may render MA ineffective. Previous analyses of the vulnerability of face recognition systems to MA, e.g., in [15], used probe images that resembled passport images and, hence do not apply to our MAD database. Therefore, we evaluated whether face recognition systems are also vulnerable to MA based on our MAD database.

The vulnerability was evaluated using the metrics presented in [25], 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. The NPP-set of the database is tested against an open-source face recognition system, namely ArcFace [43], and one commercial off-the-shelf system, referred to as COTS[6]. First, the face recognition performance of the systems is evaluated with the bona fide and probe images (resembling the eGate). The vulnerability analysis is performed exclusively on the NPP images, as preliminary studies have shown that the performance of facial recognition systems is only slightly affected by post-processing. On each database, the thresholds of the undisturbed face recognition systems are set to achieve a False Match Rate (FMR) of 0.1% according to the FRONTEX recommendation for border control scenarios [44]. The results are presented in Table III 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%. 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 the fact, that the probes of FRGCv2 pose a much higher variation in illumination, sharpness and expression. Further, it can be noted, that the impostor distributions of FERET and FRGCv2 are close to each other, thus the thresholds are approximately the same over both datasets.

Second, the vulnerability of the face recognition systems is evaluated. ArcFace and the COTS system are both very vulnerable to the MADs contained in the database. Especially morphs of the FERET dataset are almost completely accepted.

[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.

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



(a) FERET ArcFace

(b) FERET COTS

(c) FRGCv2 ArcFace

(d) FRGCv2 COTS

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. The estimated threshold for a FAR of 0.1% is depicted by τ

In general, morphs created by morphing algorithms producing less artefacts (FaceFusion and UBO) are more likely to pass the recognition system. In addition, it should be noted that the morphs generated from the FERET database are generally more successful MAs than the morphs from FRGCv2. This can be attributed to the different genuine distribution curves of the two databases. Since the genuine comparisons of the FERET data set achieve higher similarity scores, the morph comparisons also tend to do the same.

IV. MAD BASED ON DEEP FACE REPRESENTATIONS

The biggest issue regarding existing differential MAD algorithms is that they cannot cope with the large 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 such deep face recognition systems for differential MAD. Precisely, we use the following deep face recognition systems for MAD:

- ArcFace [43], an up-to-date network with a topology optimized for automatic facial recognition
- A commercial face recognition SDK from Eyedea[7]
- A re-implementation[8] of the FaceNet algorithm: [45]

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

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

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 [15], 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). While these deep features have not been trained to detect MAs, at least in the differential scenario, they can, nevertheless, be very useful for MAD: As the morphed face image does not only contain biometric information of the attacker but also those of the accomplice, its deep face features should, 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.

Most facial recognition systems work as shown in Figure 7. The facial image is normalized and transferred to the trained neural network for feature extraction. This net transfers the

TABLE III
VULNERABILITY EVALUATION OF FACE RECOGNITION SYSTEMS (FRSS)

FRS	Database	Threshold @ FMR = 0.1%	FNMR @ FMR = 0.1%	FaceFusion		FaceMorpher		OpenCV		UBO-Morpher	
				MMPMR	RMMR	MMPMR	RMMR	MMPMR	RMMR	MMPMR	RMMR
ArcFace	FERET	-1.10	0%	93.0%	93.0%	75.7%	75.7%	79.9%	79.9%	92.9%	92.9%
	FRGCv2	-1.06	0%	77.2%	77.2%	50.0%	50.0%	54.0%	54.0%	72.4%	72.4%
COTS	FERET	0.54	0%	96.0%	96.0%	88.6%	88.6%	91.3%	91.3%	96.9%	96.9%
	FRGCv2	0.58	0%	79.4%	79.6%	60.1%	60.3%	62.8%	63.0%	81.5%	81.7%

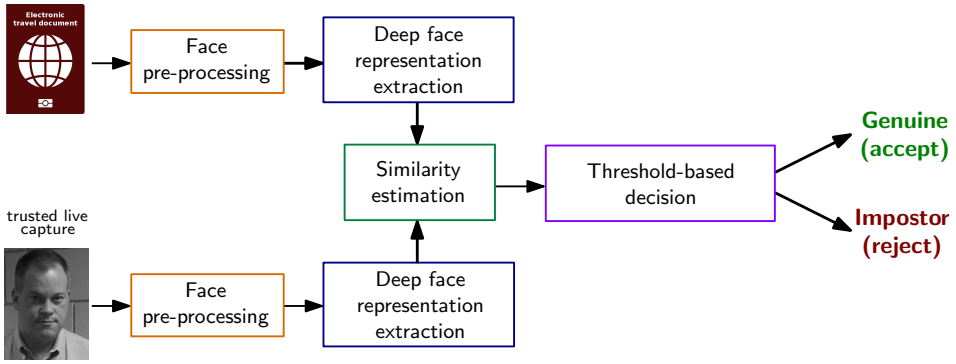


Fig. 7. Schematic of face recognition systems based on neural networks

facial image into a discriminatory feature space with smaller dimensions (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 distance measure is not suitable to separate morphs from bona fide images, the feature vectors can nevertheless contain sufficient information to detect morph attacks.

Even a simple dimension reduction of the difference of the feature vectors by Multi-Dimensional Scaling (MDS) to two dimensions shows that bona fide and morphs can be separated. Corresponding scatter plots are shown in Figure 8 and Figure 9. A plot shows data points for morphs of all four morphing algorithms. Since, as shown in Section III-C, the subset of the FRGCv2 allows more comparisons, the corresponding scatter plot (Figure 9) is much denser compared to the one generated from the FERET data. Nevertheless, for both data sets it can be seen that they can be separated on the basis of the ArcFace and Eyedea features. Especially the FERET subset can be separated almost error-free with ArcFace features. The FaceNet features also allow a separation, but at a higher error rate.

The processing of our MAD algorithms is shown in Figure 10. Both pre-processed images are fed into the neural network, the resulting deep feature vectors are combined and subsequently processed by a previously trained classifier. A simple but effective combination of the deep features is subtraction, which preserves the dimensionality of the feature vector and, thus, 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.

TABLE IV
NUMBER OF COMPARISONS PER TEST SET.

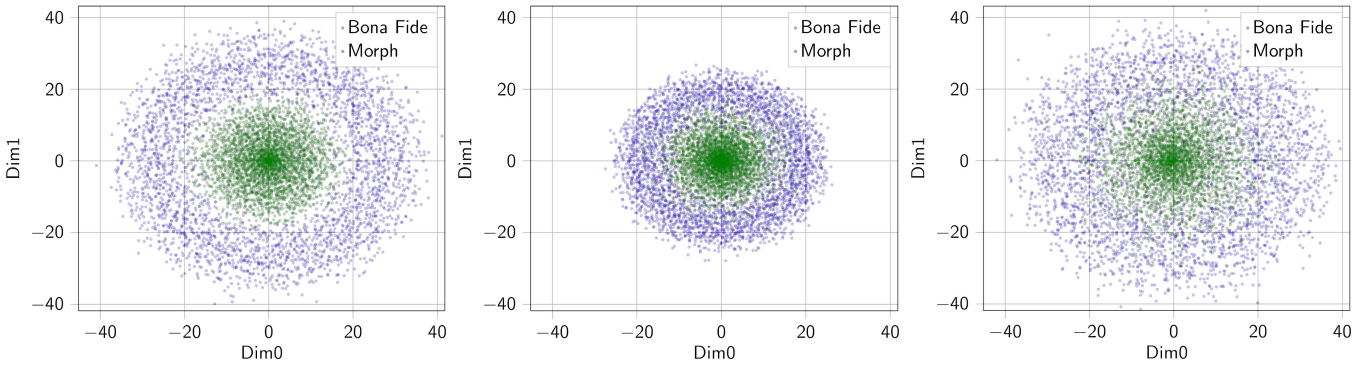
Database	Bona Fide	Morph Attacks
FERET	791	791
FRGCv2	3,298	3,246

V. EXPERIMENTS

Based on the MAD database described in Section III, we evaluate the potential of our MAD approach based on deep face representations described in Section IV. In our first experiment, we benchmark the accuracy of our approach based on deep face representations 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 in the causes of classification errors.

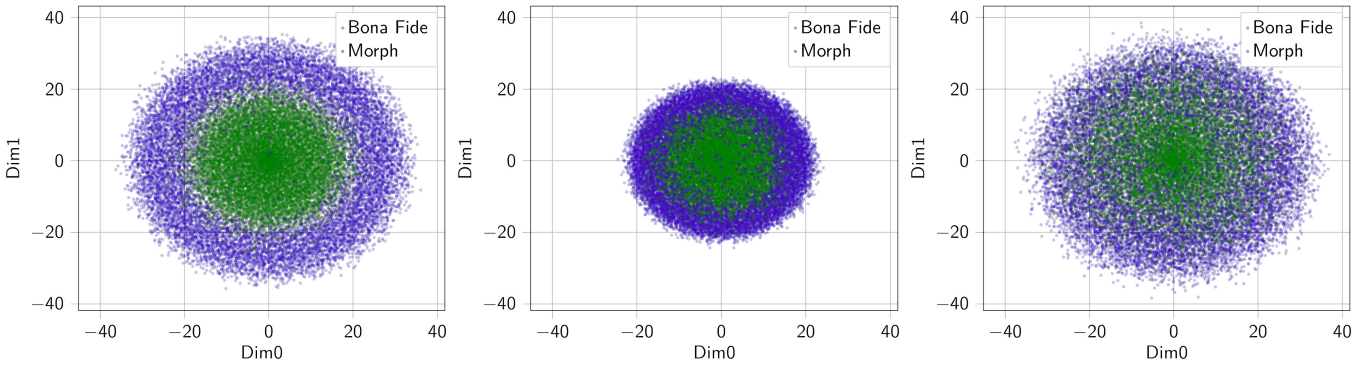
For all evaluations presented in this section, we separate training and test set by source database; precisely, all algorithms are trained using the default hyperparameters on images originating from FERET and evaluated on images from FRGCv2, or vice versa. This approach does not only ensure a strict 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 and test set 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 or test set. The numbers of comparisons performed for each test set are listed in Table IV.

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



(a) ArcFace (b) Eyedea (c) Facenet

Fig. 8. Scatter plots of MDS-reduced feature vectors generated from FERET



(a) ArcFace (b) Eyedea (c) Facenet

Fig. 9. Scatter plots of MDS-reduced feature vectors generated from FRGCv2

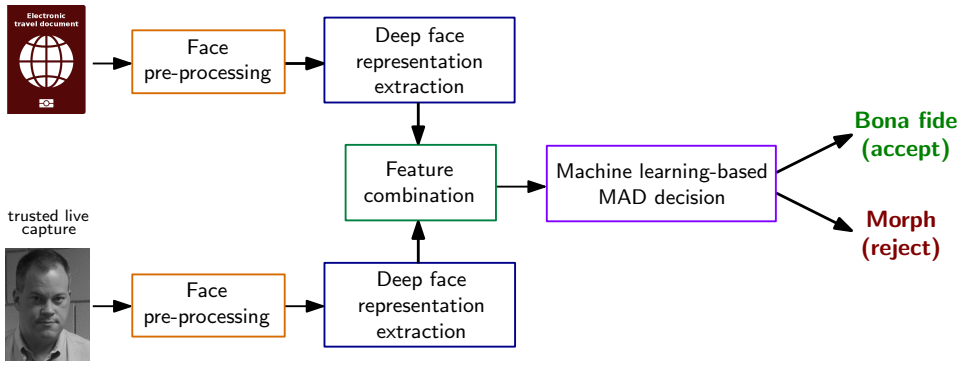


Fig. 10. Schematic of differential MAD systems based on deep face representations

classified as presentation attack (BPCER). For APCER and BPCER the definitions of ISO IEC 30107-3 [46] 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

A. Accuracy in the Absence of Post-Processing and Comparison with other MAD Algorithms

In this section, we evaluate the accuracy of our MAD approach on the NPP images and compare its accuracy to that of several other differential MAD approaches. Beside the classifiers trained on deep face features of ArcFace, Eyedea and FaceNet, we test two MAD algorithms based on texture features, namely LBP [23] and BSIF [24] with patches of size 3×3 and an optional clustering into 16 cells. In order to leverage them to differential algorithms, the difference between the histograms of reference and probe are used as features as proposed in [4]. Further, two landmark based

TABLE V
PERFORMANCE IN TERMS OF D-EER OF COMPARED ALGORITHMS ON NON-POST-PROCESSED IMAGES.

			MAD Feature Extraction									
Train DB	**Train MA**	**Test MA**	ArcFace	FaceNet	Eyedea	BSIF 1×3	BSIF 4×3	LBP 1×3	LBP 4×3	LM-Wing	LM-Dlib	De-morphing
FERET	FaceFusion	FaceFusion	**6.2%**	25.9%	15.1%	39.6%	39.8%	38.7%	41.0%	42.6%	44.6%	8.6%
		FaceMorpher	**3.6%**	21.7%	11.9%	41.9%	43.6%	38.7%	42.8%	41.5%	43.6%	3.4%
		OpenCV	**3.8%**	21.8%	12.0%	42.8%	43.5%	39.8%	42.7%	42.5%	43.9%	3.8%
		UBO-Morpher	**6.1%**	25.5%	16.1%	42.5%	43.3%	39.3%	42.8%	42.3%	44.4%	7.8%
	FaceMorpher	FaceFusion	**7.2%**	26.3%	16.0%	49.3%	46.7%	48.2%	46.3%	43.9%	43.8%	-
		FaceMorpher	**3.3%**	19.8%	9.9%	50.2%	47.6%	48.2%	45.5%	42.0%	43.1%	-
		OpenCV	**3.7%**	20.8%	10.7%	51.3%	49.2%	48.8%	47.2%	43.2%	42.8%	-
		UBO-Morpher	**6.6%**	25.2%	16.9%	51.1%	48.8%	48.1%	47.4%	43.8%	42.9%	-
	OpenCV	FaceFusion	**6.8%**	26.8%	16.3%	46.9%	43.4%	45.0%	43.8%	43.2%	45.5%	-
		FaceMorpher	**3.2%**	20.2%	10.8%	46.5%	44.9%	42.9%	43.2%	41.2%	44.3%	-
		OpenCV	**3.4%**	20.2%	11.1%	48.3%	45.8%	45.0%	44.3%	42.0%	44.4%	-
		UBO-Morpher	**6.3%**	24.9%	17.0%	48.3%	45.4%	44.8%	44.8%	42.8%	45.1%	-
	UBO-Morpher	FaceFusion	**6.7%**	26.7%	16.1%	47.0%	44.4%	43.8%	44.0%	42.8%	44.8%	-
		FaceMorpher	**3.9%**	21.9%	13.0%	45.8%	46.7%	39.8%	44.2%	41.6%	44.3%	-
		OpenCV	**3.9%**	21.7%	13.1%	47.1%	46.8%	42.7%	44.3%	42.5%	44.5%	-
		UBO-Morpher	**5.7%**	24.5%	15.3%	47.2%	45.6%	41.6%	43.1%	41.7%	44.4%	-
FRGCv2	FaceFusion	FaceFusion	**2.2%**	13.0%	4.7%	15.6%	15.9%	18.1%	25.9%	39.5%	39.0%	6.7%
		FaceMorpher	**1.2%**	9.6%	3.4%	14.6%	13.7%	12.9%	21.9%	41.3%	37.7%	4.1%
		OpenCV	**1.4%**	9.9%	3.7%	15.9%	16.7%	17.0%	24.3%	41.4%	39.5%	4.2%
		UBO-Morpher	**2.6%**	13.0%	6.6%	16.3%	18.1%	17.1%	24.3%	40.3%	39.6%	7.7%
	FaceMorpher	FaceFusion	**2.7%**	14.4%	7.2%	24.9%	23.3%	33.0%	37.7%	40.6%	39.8%	-
		FaceMorpher	**0.9%**	8.5%	3.0%	15.2%	15.6%	24.7%	31.8%	40.9%	39.8%	-
		OpenCV	**1.2%**	8.2%	3.9%	19.2%	19.4%	30.3%	34.7%	41.9%	40.9%	-
		UBO-Morpher	**2.6%**	13.7%	8.0%	21.6%	20.8%	30.4%	35.4%	41.0%	42.1%	-
	OpenCV	FaceFusion	**2.7%**	14.1%	7.1%	18.5%	19.4%	23.9%	30.5%	40.9%	38.2%	-
		FaceMorpher	**0.6%**	9.4%	3.0%	14.2%	15.3%	18.6%	25.6%	41.4%	35.7%	-
		OpenCV	**0.9%**	9.0%	3.4%	15.4%	18.1%	21.9%	28.9%	41.3%	38.1%	-
		UBO-Morpher	**2.7%**	13.2%	7.2%	17.5%	18.1%	21.8%	29.4%	41.9%	38.8%	-
	UBO-Morpher	FaceFusion	**3.2%**	13.4%	5.2%	20.5%	18.9%	21.8%	31.0%	41.1%	38.5%	-
		FaceMorpher	**0.8%**	9.9%	4.1%	13.7%	14.8%	16.8%	25.4%	43.0%	36.0%	-
		OpenCV	**1.2%**	9.5%	4.2%	16.5%	18.2%	20.3%	28.9%	42.5%	39.5%	-
		UBO-Morpher	**2.4%**	11.5%	5.4%	17.6%	18.2%	19.9%	28.5%	40.9%	38.7%	-

algorithms implemented according [8] are evaluated. One with landmarks computed with Dlib [47], the other with landmarks computed with Wing Loss [48]. Finally, we evaluate the de-morphing algorithm from [10] with a de-morphing 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 de-morphing algorithm is not trained, which is why for this MAD algorithm in the table the specification of the morphing algorithm used for training can be ignored.

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. Similarly, our results on the MAD based on texture features and landmarks are much worse than those reported in [4] and [8], where the probe images used had a higher quality. 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 yield 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 8 and 9 can be seen.

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 (FERET): the detection performance of these algorithms on probes with a more realistic variance (FRGCv2) 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, de-morphing in combination with the COTS face recognition system yields very good detection rates. In particular, for FRGCv2, it performs only slightly worse than our ArcFace-based algorithm for the higher quality morphs (FaceFusion and UBO-Morpher) and even comparably good for the lower quality morphs; for FERET, however, its performance even falls behind that of the Eyedea-based MAD algorithm.

Another pattern 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 detection.[9] Although the best results are achieved if train and test set are based on the same morphing tool (which indicates slight overfitting to the

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

characteristics of the morph), 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 w.r.t. post-processing of the reference images is analyzed. All four post-processings presented in Section III-E are used for training and testing separately. 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 except for De-Morphing. The robustness of De-Morphing on printed and scanned images was already shown in [10]. The results for morphing algorithms based on deep face representations are shown in Table VI.

As in the previous experiment, the MAD algorithm based on ArcFace-features is far ahead in detection performance, followed by the MAD algorithm based on Eyedea-features. 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, it can be said that the algorithms are not severely influenced by the tested post-processings. Presumably, this is due to the extraction of the features 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 the algorithm based on ArcFace-features 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. Figure 11 shows the 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), but evaluation was performed on the images of 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 used for the test set. 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 very low score) misclassified as attacks, visible in Figure 11 (a) to (d) as a peak for bona fides at 0. On the contrary, on FERET, fatal misclassifications occur mostly for attacks, which can be seen in Figure 11 (e) to (h) as small bumps in the morph score distribution close to 1. Examples for the above mentioned errors are given in Figs. 12 and 13. If the morph is of good quality and very similar to the sample, the MAD algorithm can no longer detect it. For example, in the samples shown in Figure 12 it is difficult even for an experienced observer to detect the morph. Examples for wrongly classified bona fide samples are given in Figure 13. The cause of this error is obvious. If the quality of the probe image is too low, it negatively affects the detection performance of the MAD system. 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.

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

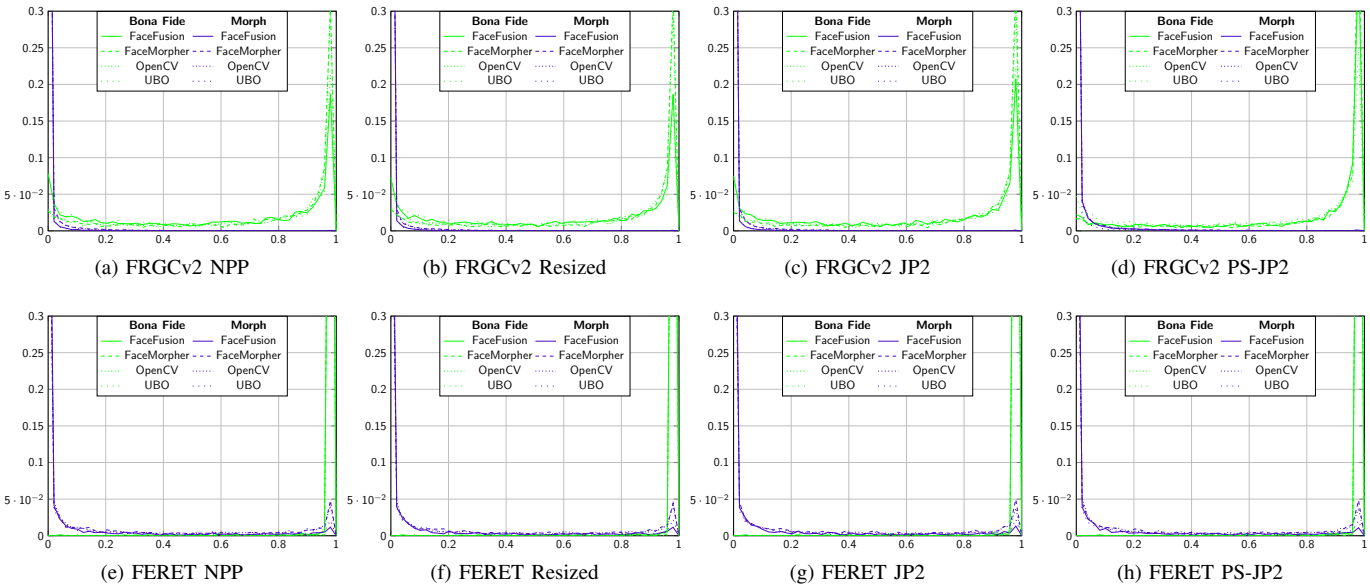
VI. Conclusion

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.

TABLE VI
DETECTION PERFORMANCE OF ALGORITHMS BASED ON DEEP FACE REPRESENTATIONS WITH POST-PROCESSINGS

				Train PP											
				NPP	Resized	JP2	PS-JP2	NPP	Resized	JP2	PS-JP2	NPP	Resized	JP2	PS-JP2
Train DB	**Train MA**	**Test MA**	**Test PP**	ArcFace				Eyedea				FaceNet			
FERET	FaceMorpher	FaceFusion	NPP	7.2%	7.2%	7.2%	7.3%	16.0%	16.1%	16.5%	16.4%	26.3%	25.9%	26.3%	26.7%
			Resized	7.2%	7.1%	7.2%	7.3%	16.0%	16.3%	16.4%	16.5%	27.4%	27.0%	27.6%	28.0%
			JP2	7.3%	7.3%	7.1%	7.3%	16.3%	16.4%	16.7%	16.6%	27.6%	27.6%	27.6%	28.1%
			PS-JP2	7.8%	7.7%	7.5%	7.8%	16.6%	16.9%	17.0%	16.4%	26.9%	26.9%	26.8%	27.8%
		UBO-Morpher	NPP	6.6%	6.7%	6.6%	6.9%	16.9%	17.1%	17.3%	17.0%	25.2%	24.7%	24.9%	25.8%
			Resized	6.7%	6.7%	6.6%	6.8%	17.0%	17.1%	17.5%	17.0%	25.6%	25.2%	25.6%	26.1%
			JP2	6.6%	6.6%	6.6%	7.0%	17.1%	17.2%	17.5%	17.2%	25.6%	25.7%	25.9%	26.2%
			PS-JP2	7.0%	7.0%	7.0%	6.7%	17.7%	17.9%	18.0%	17.4%	25.8%	26.2%	26.2%	26.8%
	OpenCV	FaceFusion	NPP	6.8%	6.7%	6.8%	7.0%	16.3%	16.5%	16.6%	16.5%	26.8%	25.5%	25.9%	26.5%
			Resized	6.8%	6.6%	6.8%	7.1%	16.3%	16.4%	16.5%	16.7%	27.2%	26.2%	26.2%	27.0%
			JP2	6.7%	6.6%	6.7%	6.8%	16.2%	16.5%	16.6%	16.6%	27.2%	26.8%	26.8%	27.3%
			PS-JP2	7.2%	7.1%	7.0%	7.1%	16.7%	16.9%	17.2%	16.8%	26.9%	26.2%	26.4%	27.1%
		UBO-Morpher	NPP	6.3%	6.3%	6.4%	6.6%	17.0%	16.9%	17.1%	16.9%	24.9%	24.2%	24.2%	25.1%
			Resized	6.4%	6.2%	6.3%	6.5%	16.8%	16.9%	17.0%	17.0%	25.7%	24.8%	24.6%	25.4%
			JP2	6.2%	6.1%	6.3%	6.5%	17.0%	17.0%	17.3%	17.0%	25.8%	25.2%	25.0%	25.5%
			PS-JP2	6.7%	6.6%	6.5%	6.5%	17.3%	17.6%	17.7%	17.3%	25.9%	25.7%	25.7%	25.8%
FRGCv2	FaceMorpher	FaceFusion	NPP	2.7%	3.0%	2.7%	2.8%	7.2%	7.2%	7.2%	7.1%	14.4%	15.3%	14.4%	15.1%
			Resized	2.8%	3.1%	3.1%	3.0%	7.1%	7.2%	7.2%	7.1%	14.4%	15.3%	15.1%	15.8%
			JP2	3.0%	3.2%	2.7%	2.8%	7.1%	6.8%	6.8%	7.3%	14.2%	14.6%	14.4%	14.9%
			PS-JP2	3.1%	3.3%	3.0%	3.0%	5.9%	6.1%	6.3%	5.8%	10.0%	10.3%	10.6%	11.0%
		UBO-Morpher	NPP	2.6%	2.8%	2.8%	2.8%	8.0%	7.8%	8.1%	8.0%	13.7%	13.7%	13.7%	14.7%
			Resized	2.8%	2.8%	2.7%	3.1%	8.1%	8.2%	7.8%	8.5%	14.2%	14.2%	14.6%	14.8%
			JP2	3.1%	3.2%	2.8%	3.1%	8.0%	8.1%	7.8%	8.5%	13.5%	14.1%	13.8%	14.6%
			PS-JP2	3.2%	3.4%	3.7%	3.5%	8.7%	9.2%	9.2%	9.0%	15.3%	15.2%	14.6%	15.3%
	OpenCV	FaceFusion	NPP	2.7%	2.7%	2.6%	2.1%	7.1%	7.0%	7.1%	7.2%	14.1%	14.9%	15.7%	15.2%
			Resized	3.0%	3.0%	2.6%	2.4%	6.8%	6.8%	6.7%	7.2%	14.9%	14.1%	14.2%	15.6%
			JP2	2.7%	3.0%	2.8%	2.8%	7.1%	6.8%	6.8%	7.2%	14.2%	13.8%	13.8%	15.1%
			PS-JP2	2.9%	3.2%	3.0%	3.0%	5.1%	5.2%	5.2%	5.4%	9.6%	9.9%	10.0%	10.5%
		UBO-Morpher	NPP	2.7%	2.8%	2.3%	2.7%	7.2%	7.8%	7.5%	7.8%	13.2%	13.7%	13.4%	14.8%
			Resized	2.6%	2.7%	2.6%	2.7%	7.6%	7.5%	7.6%	7.8%	13.5%	13.0%	13.7%	14.8%
			JP2	2.7%	2.8%	2.7%	2.8%	7.7%	7.7%	7.3%	7.5%	13.2%	13.0%	13.5%	14.3%
			PS-JP2	3.1%	3.0%	3.0%	3.2%	8.2%	8.1%	8.0%	8.1%	14.2%	14.3%	14.4%	14.3%

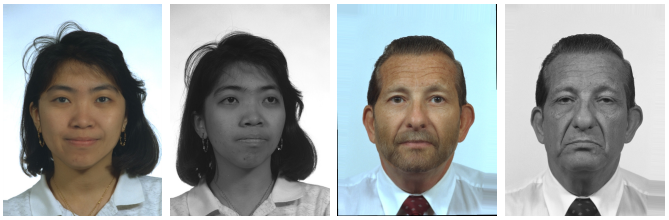


(a) FRGCv2 NPP (b) FRGCv2 Resized (c) FRGCv2 JP2 (d) FRGCv2 PS-JP2

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

Fig. 11. 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.

- *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,



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

Fig. 12. Attack presentation classification error examples on FERET

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

Fig. 13. Bona fide presentation classification error examples on FRGC

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

This work was partially supported by the German Federal Ministry of Education and Research (BMBF), by the Hessen State Ministry for Higher Education, Research and the Arts (HMWK) within the Center for Research in Security and Privacy (CRISP) as well as by the Federal Office of Information Security (BSI) within the FACETRUST project. The UBO-Morpher as well as the De-Morphing implementation was kindly provided by the Biometric System Laboratory of 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. 23 012–23 026, 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*. Springer International Publishing, 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," *Cognitive Research: Principles and Implications*, vol. 3, no. 1, jun 2018.

[4] U. Scherhag, C. Rathgeb, and C. Busch, "Towards detection of morphed face images in electronic travel documents," in *Proceedings of the 13th IAPR Workshop on Document Analysis Systems (DAS)*, 2018.

[5] P. Phillips, P. Flynn, T. Scruggs, K. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2005.

[6] U. Scherhag, D. Budhrani, M. Gomez-Barrero, and C. Busch, "Detecting morphed face images using facial landmarks," in *Proceedings of the 2018 International Conference on Image and Signal Processing (ICISP)*. Springer International Publishing, 2018, pp. 444–452.

[7] A. Martinez and R. Benavente, "The AR face database," Computer Vision Center (CVC), Tech. Rep. 24, Jun. 1998.

[8] N. Damer, V. Boller, Y. Wainakh, F. Boutros, P. Terhrst, A. Braun, and A. Kuijper, "Detecting face morphing attacks by analyzing the directed distances of facial landmarks shifts," in *Proceedings of the 40th German Conference of Pattern Recognition (GCPR)*, 2018.

[9] P. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The FERET database and evaluation procedure for face-recognition algorithms," *Image and Vision Computing*, vol. 16, no. 5, pp. 295–306, apr 1998.

[10] M. Ferrara, A. Franco, and D. Maltoni, "Face demorphing," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 4, pp. 1008–1017, apr 2018.

[11] ——, "Face demorphing in the presence of facial appearance variations," in *Proceedings of the 26th European Signal Processing Conference (EUSIPCO)*, 2018.

[12] W. Gao, B. Cao, S. Shan, D. Zhou, X. Zhang, and D. Zhao, "The CAS-PEAL large-scale chinese face database and baseline evaluations," Chinese Academy of Sciences, Tech. Rep. JDL-TR-04-FR-001, May 2004.

[13] 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. 75 122–75 131, 2019.

[14] A. Makrushin, T. Neubert, and J. Dittmann, "Automatic generation and detection of visually faultless facial morphs," in *Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*. SCITEPRESS - Science and Technology Publications, 2017.

[15] M. Ferrara, A. Franco, and D. Maltoni, "Face morphing detection in the presence of printing/scanning and heterogeneous image sources."

[16] R. Ramachandra, K. B. Raja, S. Venkatesh, and C. Busch, "Transferable deep-CNN features for detecting digital and print-scanned morphed face images," in *Proceedings of the 2017 Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, jul 2017.

[17] R. Ramachandra, S. Venkatesh, K. Raja, and C. Busch, "Towards making morphing attack detection robust using hybrid scale-space colour texture features," in *Proceedings of 5th International Conference on Identity, Security and Behaviour Analysis (ISBA 2019)*, 2019, pp. 22–24.

[18] U. Scherhag, R. Ramachandra, K. B. Raja, M. Gomez-Barrero, C. Rathgeb, and C. Busch, "On the vulnerability of face recognition systems towards morphed face attacks," in *Proceedings of the 5th International Workshop on Biometrics and Forensics (IWBF)*. IEEE, Apr. 2017.

[19] M. Ngan, P. Grother, and K. Hanaoka, "Face recognition vendor test (frvt) morph performance of automated face morph detection," National Institute of Technology (NIST), Tech. Rep. Draft NISTIR, 2019.

[20] R. Ramachandra, K. B. Raja, and C. Busch, "Detecting morphed face images," in *Proceedings of the 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. IEEE, sep 2016.

[21] 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 *Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security - IHMMSec '17*. ACM Press, 2017.

[22] C. Seibold, W. Samek, A. Hilsmann, and P. Eisert, "Detection of face morphing attacks by deep learning," in *Digital Forensics and Watermarking*. Springer International Publishing, 2017, pp. 107–120.

[23] T. Ojala, M. Pietikinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, jan 1996.

[24] J. Kannala and E. Rahtu, "BSIF: Binarized statistical image features," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, Nov 2012, pp. 1363–1366.

[25] U. Scherhag, A. Nautsch, C. Rathgeb, M. Gomez-Barrero, R. N. J. Veldhuis, L. Spreeuwers, M. Schils, D. Maltoni, P. Grother, S. Marcel, R. Breithaupt, R. Ramachandra, and C. Busch, "Biometric systems under morphing attacks: Assessment of morphing techniques and vulnerability reporting," in *Proceedings of the 2017 International Conference of the Biometrics Special Interest Group (BIOSIG)*. IEEE, sep 2017.

[26] L. Spreeuwers, M. Schils, and R. Veldhuis, "Towards robust evaluation of face morphing detection," in *Proceedings of the 26th European Signal Processing Conference (EUSIPCO)*, 2018.

[27] 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 *Proceedings of the 9th IEEE International Conference on Biometrics: Theory, Applications, and Systems (BTAS)*. IEEE, 2018.

[28] R. Ramachandra, K. Raja, S. Venkatesh, and C. Busch, "Face morphing versus face averaging: Vulnerability and detection," in *Proceedings of the 2017 International Joint Conference on Biometrics (IJCB)*. IEEE, oct 2017.

[29] L. Wandzik, G. Kaeding, and R. V. Garcia, "Morphing detection using a general-purpose face recognition system," in *Proceedings of the 26th European Signal Processing Conference (EUSIPCO)*, 2018.

[30] A. Asaad and S. Jassim, "Topological data analysis for image tampering detection," in *Digital Forensics and Watermarking*. Springer International Publishing, 2017, pp. 136–146.

[31] S. Jassim and A. Asaad, "Automatic detection of image morphing by topology-based analysis," in *Proceedings of the 26th European Signal Processing Conference (EUSIPCO)*, 2018.

[32] A. Agarwal, R. Singh, M. Vatsa, and A. Noore, "SWAPPED! digital face presentation attack detection via weighted local magnitude pattern," in *Proceedings of the 2017 International Joint Conference on Biometrics (IJCB)*. IEEE, oct 2017.

[33] L. Debiasi, U. Scherhag, C. Rathgeb, A. Uhl, and C. Busch, "PRNU-based detection of morphed face images," in *Proceedings of the 6th International Workshop on Biometrics and Forensics (IWBF)*. IEEE, 2018.

[34] L. Debiasi, C. Rathgeb, U. Scherhag, A. Uhl, and C. Busch, "PRNU variance analysis for morphed face image detection," in *Proceedings of the 9th IEEE International Conference on Biometrics: Theory, Applications, and Systems (BTAS)*. IEEE, 2018.

[35] U. Scherhag, L. Debiasi, C. Rathgeb, C. Busch, and A. Uhl, "Detection of face morphing attacks based on PRNU analysis," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, pp. 1–1, 2019.

[36] L. Debiasi, N. Damer, A. M. Saladi, C. Rathgeb, U. Scherhag, C. Busch, F. Kirchbuchner, and A. Uhl, "On the detection of gan-based face morphs using established morph detectors," in *Proceedings of the 20th International Conference on Image Analysis and Processing (ICIAP)*, 2019.

[37] L.-B. Zhang, F. Peng, and M. Long, "Face morphing detection using fourier spectrum of sensor pattern noise," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, jul 2018.

[38] 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 *Proceedings of the 5th International Workshop on Biometrics and Forensics (IWBF)*. IEEE, apr 2017.

[39] C. Seibold, A. Hilsmann, and P. Eisert, "Reflection analysis for face morphing attack detection," in *Proceedings of the 26th European Signal Processing Conference (EUSIPCO)*, 2018.

[40] C. Seibold, W. Samek, A. Hilsmann, and P. Eisert, "Accurate and robust neural networks for security related applications exampled by face morphing attacks," *Computer Vision and Pattern Recognition*, pp. 1–16, 2018.

[41] International Civil Aviation Organization, "ICAO doc 9303, machine readable travel documents – part 9: Deployment of biometric identification and electronic storage of data in MRTDs (7th edition)," ICAO, Tech. Rep., 2015.

[42] European Commission, "EU-eMRTD specification," European Commission, Tech. Rep., 2018.

[43] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," 2018.

[44] Research and Development Unit, "Best practice technical guidelines for automated border control (abc) systems," FRONTEX, Tech. Rep., 2012.

[45] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proceedings of the 2015 Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2015.

[46] ISO/IEC JTC1 SC37 Biometrics, "Information technology – biometric presentation attack detection – part 3: Testing and reporting," International Organization for Standardization, Geneva, Switzerland, ISO ISO/IEC IS 30107-3:2017, 2017.

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

[48] Z.-H. Feng, J. Kittler, M. Awais, P. Huber, and X.-J. Wu, "Wing loss for robust facial landmark localisation with convolutional neural networks."