Morphing Attack Detection: A Fusion Approach

Siri Lorenz, Ulrich Scherhag, Christian Rathgeb and Christoph Busch
da/sec – Biometrics and Internet Security Research Group, Hochschule Darmstadt, Germany

{siri.lorenz,christian.rathgeb,christoph.busch}@h-da.de, ulrich.scherhag@icognize.de

***Abstract*—Face morphing attacks pose a serious threat to existing face recognition systems. As a number of studies have shown, existing face recognition systems and human experts can be fooled by morphed facial images. Based on these findings various approaches to morphing attack detection have been published. Automated morphing attack detection is still a young branch of research with many recent publications.**

Using features extracted by different feature extractors we develop a score-level based fusion approach. The scores are generated by different classifiers with optimised hyperparameters. We use different approaches to determine the weights for the sum-rule: grid-search and random forests scoring function as well as normalised scores.

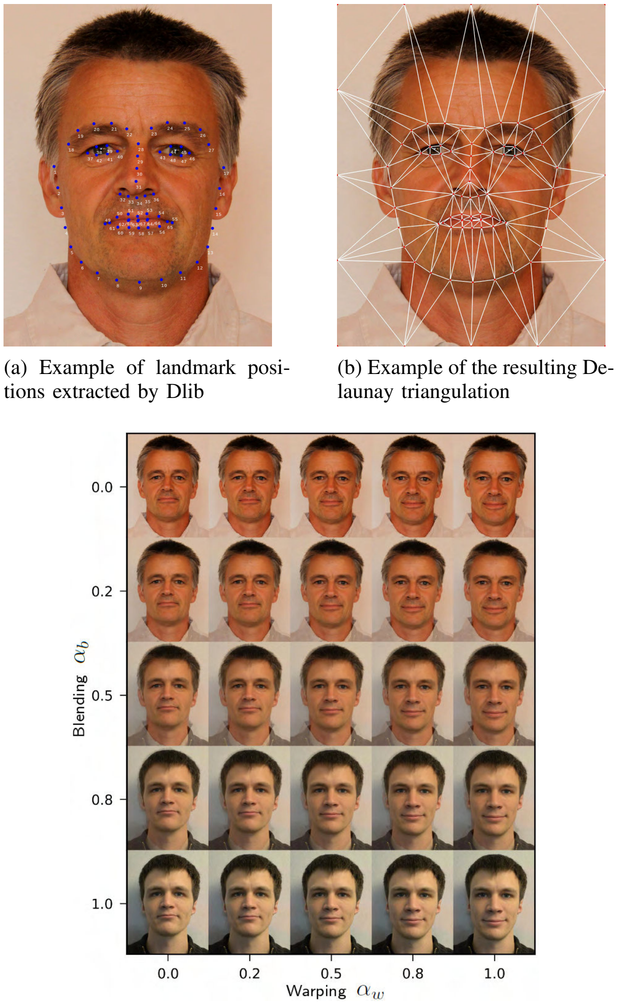
We notice that a weighted score-level fusion can achieve improved results. Moreover, we observe that weights determined by grid-search might lead to better results when using fewer scores compared to those obtained by random forest while the former is more time consuming. However, both random forest and grid-search weights can significantly improve the morphing attack detection performance.

***Index Terms*—Face Morphing, Hyperparameter Optimisation, Morphing Attack Detection, Biometric Fusion**

I. Introduction

Face Morphing poses a serious threat to existing face recognition systems which are used in many application scenarios - in particular in Automated Border Control (ABC). As first shown in [1], two individuals can reach a match decision when being compared to a single morphed reference created from facial images of these individuals. Using this method ABC systems as well border control guards can be circumvented by using a passport containing a morphed image.

Morphing is the targeted combination of two or more (facial) images, where the correspondence between striking characteristics (so called landmarks) is determined. These corresponding landmarks are geometrically aligned e.g. using Delaunay triangulation. This results in a distortion of the image - this is referred to as warping - and the landmarks of each individual can be weighted by an α-value (α_w). Lastly the texture values of both images are blended - the texture values can also be weighted by an α-value (α_b) [2]. Examples for landmarks (Figure 1a) and the resulting Delaunay triangulation (Figure 1b) as well as examples for different warping and blending values (Figure 1c) can be seen in Figure 1.



(a) Example of landmark positions extracted by Dlib

(b) Example of the resulting Delaunay triangulation

(c) Example for warping and blending α-value

Fig. 1: Example of a morphing process: (a) landmark detection, (b) triangulation, (c) warping and blending with different weights between two facial image (From: [2])

This research work has been funded by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

Since the first publication various approaches for Morphing Attack Detection (MAD) have been suggested, however the detection performance of these approaches is still improvable. Therefore, this paper will evaluate the results of a score-level fusion approach to MAD. Our approach uses score-level fusion and different algorithms to determine the weights for the fusion. In addition the models were trained using optimised hyperparameters.

This paper is structured as follows. In section II we will give an overview of related work and used metrics. In section III the dataset will be briefly introduced and in section IV the fusion approaches will be explained. The results of our evaluation will be presented in section V. Finally our conclusions are drawn in section VI and future work is discussed in section VII.

II. Background

This section gives an outline of used metrics, the basic concept of MAD, and related work.

A. Metrics

The following metrics are used in this work:

Attack Presentation Classification Error Rate: The Attack Presentation Classification Error Rate (APCER) is the proportion of attack presentations classified as bona fide presentations [3].

BonaFide Presentation Classification Error Rate: The Bonafide Presentation Classification Error Rate (BPCER) is the proportion of bona fide presentations classified as attack presentations [3].

BPCER10: Is the BPCER where the APCER is 10% [3].

BPCER20: Is the BPCER where the APCER is 5% [3].

Detection Equal Error Rate: The Detection Equal Error Rate (D-EER) is the operating point where APCER and BPCER are equal [4].

B. Morphing Attack Detection

MAD can be divided into two categories:

Single image MAD: If only the suspected face image is available, single image MAD (S-MAD) can be applied. In this scenario only a single image is parsed and evaluated. One example for S-MAD is the passport application process, when a printed face image is submitted and scanned.

Differential MAD: For differential MAD (D-MAD) the suspected image and a trustworthy image (e.g. a trusted live capture) are compared to each other. D-MAD could for example be applied in an ABC scenario [2].

Most S-MAD algorithms can be extended to D-MAD algorithms. However, some MAD algorithms need a trusted live capture and can therefore only be used as D-MAD. One of which is a landmark-based approach. For landmark-based approaches the absolute coordinates of the landmarks from both images are normalised and the Euclidean distance is computed or the angle between the landmarks is used to counter the variations of two images in expressions or pose. For this the angle is normalised to a range of 0° to 180° [5].

Some approaches are more suitable for S-MAD while various approaches can be used for S-MAD and D-MAD some of which were used in our experiments.

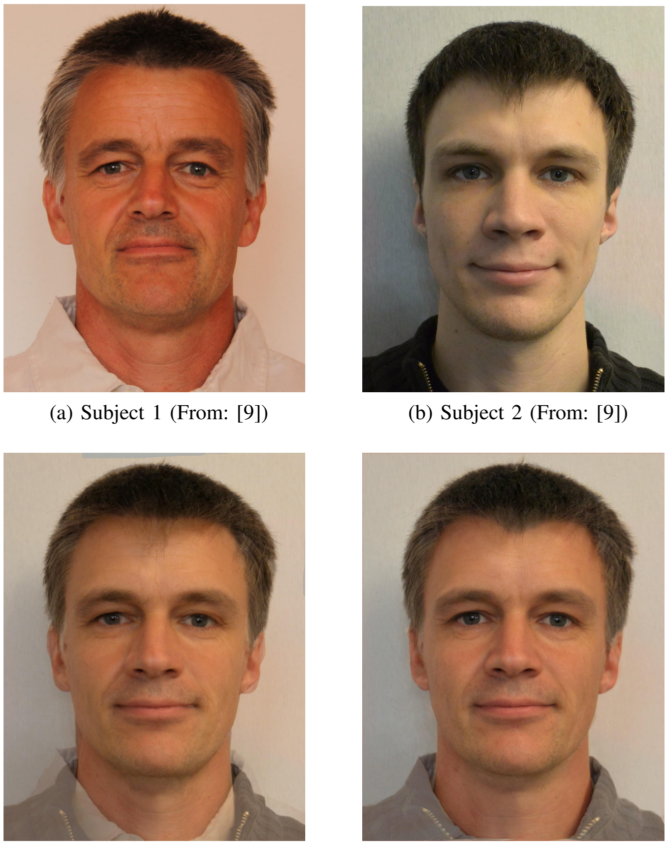
1) Image Noise Pattern: Image noise pattern approaches are more suitable for S-MAD than for D-MAD. Images taken with a modern digital camera have certain noise pattern. These patterns can be extracted and analysed. If, however, a morphed image is analysed, then a combination of at least two patterns will be found. Due to the alignment of the landmarks during the morphing process, the patterns can also be distorted and therefore altered. One approach [6] related to image noise patterns is Photo Response Non-Uniformity (PRNU) analysis, where the unique PRNU-pattern is extracted and analysed [6]–[8]. In their approach [6] the authors conclude, that although their approach is robust against image scaling and sharpening, it is not robust against histogram equalisation and further research is needed.

2) Texture Descriptors: Texture descriptors are suitable for S-MAD as well as D-MAD. Prominent examples for texture descriptors are Local Binary Pattern (LBP) and Binarized Statistical Image Features (BSIF). Specific artefacts may result from the morphing process. Texture descriptors can be used to detect these artefacts. The most common artefacts occur in the region of the eyes and ghost artefacts surrounding the neck and hair (for automated morphs). Ghost artefacts are the results of too few landmarks in a region. As a results the blending and warping process is imprecise and results in blurry regions [9]. A comparison between an automatically and manually morphed image is shown in Figure 2.

3) Deep Features: Deep Features like texture descriptors are suitable for S-MAD and D-MAD. In addition to texture descriptors or image noise pattern, machine learning can be used to extract meaningful features from a facial image. The difficulty lies within the training of the feature extractor to avoid overfitting. However, a number of pre-trained models for face recognition are available, like ArcFace and FaceNet. These convolutional neural networks compute a feature vector which can be analysed for MAD [10], [11].

4) Generative Adversarial Nets: Generative Adversarial Nets (GANs) are a machine learning approach where two models are used to train each other. These two models are not working together but against each other, hence the name *adversarial*. One of the models is called generative model - it generates morphed images from non-morphed images - the other is called discriminative model - it is trained to estimate whether an image is generated or a bona fide image. The estimations are returned to the generative model which uses the predictions to generate more realistic morphs [12]. With every iteration the morphed images become more realistic and harder to detect, but, as a result, the discriminative model also evolves and becomes better in differentiating between morphed and non-morphed images [13].

GANs can be used used for MAD in various ways, one of which is a face de-morphing GAN-based approach introduced in [14]. Face de-morphing subtracts the trusted live capture from the suspected image by applying a reversed morphing



(a) Subject 1 (From: [9])

(b) Subject 2 (From: [9])

(c) Automatically morphed - showing artefacts around the iris, hair, and neck (From: [9])

(d) Manually morphed - showing no (human visible) artefacts (From: [9])

Fig. 2: Comparison of a manual and automated morph (From: [9])

operation. The resulting image is then compared to the live capture. If it was a morphed image the accomplice should remain, otherwise the subject should remain and a verdict can be reached based on the result. The authors of [14] introduce a face de-morphing GAN (FD-GAN). Although there are still unsolved problems, the FD-GAN can achieve good results in the author's experimental environment [14].

C. Related Work

The authors of [15] published a multi-algorithm fusion approach to S-MAD using OpenCV to generate the morphed images.

The authors of [16] introduce a method to adopt the Dempster-Shafer theory to MAD. The authors combine various individual MAD methods using Dempster-Shafer theory. Their experiments show that their fusion approach improves the accuracy of the results significantly, compared to the individual detectors and also compared to fusion approaches using majority voting or an unweighted sum-rule [16].

The authors of [17] point out the difficulty of S-MAD compared to D-MAD, as all information need to be taken from one image without any reference and introduce an approach to improve S-MAD using an ensemble of features. Their experiments indicate a significant improvement in robustness and reliability, especially to different types of printers on their datasets [17].

Unlike the authors of [15] and those of [17] we use S-MAD and D-MAD, as well as morphed facial images created using FaceFusion[1], FaceMorpher[2], a morphing tool of the University of Bologna, and OpenCV[3] [18]. In comparison to the authors of [16] we combine scores using the sum-rule, however, we also reach the conclusion that the unweighted sum-rule achieves weak results compared to a weighted sum-rule.

The authors of [19] explain various weighting approaches for multi-biometric-fusion and propose two additional approaches. We use three different approaches for weighting using equal weights, brute force weights and the feature importance decided by random forest. These approaches will be discussed in more detail in section IV.

The authors of [20] focus on decreasing the BPCER. They also address the low generalisation in MAD when using different morphing techniques.

III. Dataset

The morphing database used for this work is based on two databases - FERET and FRGCv2 [21], [22] - which were extended by morphed facial images. All images were post-processed in various ways to ensure a diverse dataset and the morphed images were created using four different morphing algorithms (FaceFusion, FaceMorpher, OpenCV, and UBO-MA[4]). For a more detailed description of the database we refer to [18]. Models trained on FERET were evaluated on FRGCv2 and vice versa. All models were trained on unprocessed images, to evaluate the effect that post-processing has on MAD. For classification four different algorithms were used - Support Vector Machines (SVMs), Random Forest, AdaBoost, and GradientBoosting - whose hyperparameters were tuned in three different ways: using grid-search, a genetic algorithm, and Bayesian Optimisation. For our fusion we chose twenty models evaluated on sixteen different test sets each resulting in 320 scores for each database. We used these conventional classifiers as they are easier to train than neural net classifier.

IV. Fusion Approach

This section will give a short introduction to biometric fusion applicable to the context of MAD and explain our fusion approach.

A. Biometric Fusion

There are various approaches to biometric fusion which are described in detail in [23], [24]. Mainly three levels of fusion are mentioned: feature-level fusion, score-level fusion

[1] www.wearemoment.com/FaceFusion/

[2] github.com/alyssaq/facemorpher

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

[4] The morphing tool provided by the University of Bologna.

and decision-level fusion. In our approach, we used a multi-algorithm score-level-based approach. We varied the amount of algorithms used in the fusion, using two, three, four, five, fifteen or twenty scores. For every trained model we got sixteen scores, as we tested each model on all post-processing methods and every morphing algorithm. For example a model trained on FERET with differential features extracted by ArcFace using no post-processing and morphs generated by OpenCV would be evaluated on FRGCv2 with differential features extracted by ArcFace and all four post-processing methods and all four morphing tools. Out of all these scores the twenty overall best scores were chosen for each morphing tool. This included some scores from overfitted models which we decided to use in order to test the weighting of the following sum-rule implementations.

Our fusion approach was implemented based on the sum-rule:

$$s = \sum_{i} w_i s_i \tag{1}$$

where the w_i is the weight for the i-th score s_i.

B. Score-Level Fusion

In a first approach we decided to use a normalised score over all scores in the fusion-group. We tested all possible combinations of the best twenty scores of each morphing tool. While some unweighted setups - mostly those consisting of Deep Features like ArcFace, FaceNet, or Eyedea and Texture Descriptors like BSIF and LBP - could result in an improvement of the D-EER, most unweighted setups resulted in a deterioration.

C. Weighted Score-Level Fusion

To achieve better results we decided to adjust the weights for the scores using two different approaches: grid-search and random forest.

1) Grid-Search: We implemented a grid-search trying different weights between zero and one for each score under the condition that the sum of all weights equals one. For each weight combination a weighted-sum score is evaluated and the best weights are returned. However, the runtime increased exponentially - as the grid grew exponentially - and the approach was only used for fusion-groups up to size four.

2) Random Forest: The other weighted approach was implemented using random forest. Here the single scores are used to create a matrix which is then used to fit an sklearn RandomForestClassifier model. Afterwards the *feature importance* of that classifier is used to determine the weight of each score. Using the determined weights a weighted score for the fusion-group is calculated.

D. Creation of Fusion Groups

We started with fusing groups of size two up to size twenty. Two as it was the minimum and twenty as it was the maximum. Then we chose fifteen to have an bigger fusion-group but lower than twenty. The sizes three, four and five were chosen to see if an steady improvement could be observed by increasing the size one by one.

TABLE I: Abbreviations

	Name	Abbreviation
Post-processing	no post-processing	N
	resized	R
	print/scan	P
Feature Extractor	ArcFace	A
	Eyedea	E
	LBP	L
	LM-Wing	W
	FaceNet	F
	BSIF	B
Optimisation Algorithm	grid-search	1
	genetic algorithm	2
	bayesian optimisation	3
	no optimisation	4
Morphing Algorithm	FaceFusion	f
	FaceMorpher	m
	OpenCV	o
	UBO-MA	u
Classifier	SVM	s
	Random Forest	r
	AdaBoost	a
	GradientBoosting	g
MAD-Type	single MAD	S
	differential MAD	D

We used six different feature extractors: (1) LM-Wing - which is a landmark-based feature extractor and therefore only suitable for D-MAD, (2) ArcFace, Eyedea, and FaceNet - which are deep feature based and are suitable for S-MAD and D-MAD, and (3) LBP and BSIF - which are texture descriptors and are also used for S-MAD and D-MAD.

The most interesting results were achieved by training on FRGCv2 and evaluating on FERET. Due to the highly complex combinations of hyperparameter optimisation algorithms, feature extractor and morphing algorithm for each training and testing set, we used a combination of case-sensitive one-letter abbreviations - which can be seen in Table I - to refer to the scores. In Table II the setups for the most promising fusion-groups of size four are displayed. Each setup contains four scores, and each of those scores consist of a feature extractor, a post-processing method and a morphing algorithm used on the training database and the same for the testing database, a optimisation algorithm, a classifier and the MAD-type. The most promising setups for fusion-groups of size fifteen can be seen in Table III. Each column contains one setup. Each setup is composed of four or fifteen scores. Every score gives information on 1) the used feature extractor, 2) the post-processing method and morphing tool used for training, 3) the post-processing method and morphing tool used for testing, 4) the optimisation algorithm and the used classifier, and 5) the MAD-type.

V. Evaluation

For our evaluation we separately assessed the different fusion-group sizes focussing on the five most promising results. We used the smallest D-EER of the scores in the fusion-group as reference for the improvement and calculated the relative improvement (or deterioration) in percent as shown

TABLE II: Five setups achieving the best results for a fusion of four scores (weighted with grid-search).

	Setup 4-0	Setup 4-1	Setup 4-2	Setup 4-3	Setup 4-4
Weighted	Grid	Grid	Grid	Grid	Grid
Score 1	E N-f N-m 1-r D	E N-f N-m 1-r D	E N-f N-m 1-r D	A N-o R-o 4-s D	A N-o N-m 2-a D
Score 2	F N-f N-m 1-r D	F N-u N-m 3-s D	F N-f N-m 2-g D	A N-o R-o 2-a D	E N-o N-m 1-a D
Score 3	L N-m N-m 3-g D	L N-m N-m 3-g D	L N-m N-m 3-g D	B N-u R-o 3-g D	B N-u N-m 3-g D
Score 4	W N-o N-m 3-g D	W N-o N-m 3-g D	W N-o N-m 3-g D	L N-m R-o 3-g S	L N-m N-m 3-g S

in Equation (2). We decided to use the D-EER instead of focusing on APCER or BPCER to generate an upper border for the error rates. Since the D-EER is the operating point where APCER and BPCER are equal, both APCER and BPCER can be optimised below the D-EER value.

$$diff = \frac{EER_{group}}{EER_{min}} - 1 \tag{2}$$

If the best D-EER of the group was 0.90% and the group D-EER was 0.26% the relative improvement would be -71.43% - which is the best observed improvement for our experiments.

Some of the results for our random forest weights showed anomalies, as weighting both scores in a fusion-group of size two with zero. Resulting in an D-EER of 50%. As no real insight could be gained from these results, we decided to drop these results from our analysis. Furthermore, weighting with random forest sometimes resulted in weighting the worst model the highest, since random forest uses correlation to determine the weights. If a model overfitted during training the D-EER can be above 95% which would result in a very high correlation - however reversed.

For fusion-groups up to size four a steady improvement can be seen for grid-search. Fusion-groups of size fifteen weighted using random forest achieve almost the same improvement as fusion-groups of size four using grid-search to determine the weights. For fusion-groups of size twenty no improvement could be observed.

The most promising results are achieved by fusion-groups of size four when using grid-search weights and fusion-groups of size fifteen when using random forest to determine the weights. They mostly consist of scores from texture descriptors and deep features.

A. Fusion-Groups of Size Four

The weights for the most promising fusion-groups of size four are determined using grid-search. Since, grid-search is an exhaustive search it can become very time-consuming. As

TABLE III: Five setups achieving the best results for a fusion of fifteen scores (weighted with random forest).

	Setup 15-0	Setup 15-1	Setup 15-2	Setup 15-3	Setup 15-4
Weighted	RF	RF	RF	RF	RF
Score 1	A N-o N-o 4-s D	A N-o N-o 4-s D	A N-o N-o 4-s D	A N-o N-o 4-s D	A N-o N-o 4 -s D
Score 2	A N-o N-o 1-g D	A N-o N-o 1-r D	A N-o N-o 1-g D	A N-o N-o 1-g D	A N-o N-o 1-g D
Score 3	A N-o N-o 1-r D	A N-o N-o 2-a D	A N-o N-o 1-r D	A N-o N-o 1-r D	A N-o N-o 1-r D
Score 4	A N-o N-o 2-a D	E N-f N-o 1-r D	A N-o N-o 2-a D	A N-o N-o 2-a D	A N-o N-o 2-a D
Score 5	E N-u N-o 4-s D	E N-o N-o 1-a D	E N-u N-o 4-s D	E N-u N-o 4-s D	E N-u N-o 4-s D
Score 6	L N-o N-o 3-r D	W N-u N-o 1-s D	E N-f N-o 1-r D	E N-f N-o 1-r D	L N-o N-o 3-r D
Score 7	E N-f N-o 1-r D	F N-f N-o 1-r D	E N-o N-o 1-a D	E N-o N-o 1-a D	E N-f N-o 1-r D
Score 8	E N-o N-o 1-a D	W N-o N-o 1-a D	F N-f N-o 1-r D	F N-f N-o 1-r D	E N-o N-o 1-a D
Score 9	W N-u N-o 1-s D	F N-u N-o 3-s D	W N-o N-o 1-a D	W N-o N-o 1-a D	W N-u N-o 1-s D
Score 10	F N-f N-o 1-r D	W N-m N-o 3-r D	F N-u N-o 3-s D	F N-u N-o 3-s D	F N-f N-o 1-r D
Score 11	W N-o N-o 1-a D	B N-u N-o 3-g D	B N-u N-o 3-g D	W N-m N-o 3-r D	W N-o N-o 1-a D
Score 12	F N-u N-o 3-s D	F N-f N-o 1-a D	F N-f N-o 1-a D	B N-u N-o 3-g D	F N-u N-o 3-s D
Score 13	L N-m N-o 1-s S	L N-m N-o 1-s S	L N-m N-o 1-s S	L N-m N-o 1-s S	F N-f N-o 1-a D
Score 14	L N-m N-o 3-g S	L N-m N-o 3-g S	L N-m N-o 3-g S	L N-m N-o 3-g S	L N-m N-o 1-s S
Score 15	W N-o N-o 3-g D	W N-o N-o 3-g D	W N-o N-o 3-g D	W N-o N-o 3-g D	L N-m N-o 3-g S

TABLE IV: The results achieved by the five setups for a fusion of four scores.

Group	D-EER	BPCER10	BPCER20	min. D-EER	Diff (%)
Setup 4-3	**0.26%**	0.00%	0.13%	0.90%	**-71.43%**
Setup 4-4	0.52%	0.13%	0.26%	1.80%	-71.43%
Setup 4-2	1.68%	0.52%	0.90%	5.57%	-69.92%
Setup 4-0	1.80%	0.39%	1.16%	5.57%	-67.61%
Setup 4-1	1.80%	0.26%	1.55%	5.57%	-67.61%

TABLE V: The results achieved by the five setups for a fusion of fifteen scores.

Group	D-EER	BPCER10	BPCER20	min. D-EER	Diff (%)
Setup 15-2	**0.26%**	0.00%	0.13%	0.90%	**-71.43%**
Setup 15-1	**0.26%**	0.00%	0.13%	0.90%	**-71.43%**
Setup 15-4	0.39%	0.00%	0.13%	0.90%	-57.14%
Setup 15-0	0.39%	0.00%	0.13%	0.90%	-57.14%
Setup 15-3	0.39%	0.00%	0.13%	0.90%	-57.14%

the amount of combinations and, therefore, also the amount of time needed for each grid grows exponentially with ever added score we used grid-search only up to fusions groups of size four. All Setups consist of at least two deep features and one texture descriptor. While the two most promising results contain only deep features and texture descriptors, promising results can also be achieved when including a landmark-based feature extractor. As can be seen in Table IV all D-EERs are below 2%. However, the two most promising results are below 1%. Another notable difference between these two setups and the other setups is that the most promising two fuse scores of D-MAD and S-MAD, while the other setups use only D-MAD scores.

It is notable, that the best two fusion-groups consist of only deep features and texture descriptors and use models trained on different morphing algorithms and use S-MAD and D-MAD to achieve their results. In addition, all of these groups use grid-search to determine their weights. Setup 4-0, 4-1, and 4-2 only differ by one score the results are similar. Setup 4-3, and 4-4 differ by two models and although the difference between the smallest D-EER of the single algorithms and the achieved fusion D-EER is the same, the D-EER for Setup 4-3 is twice as good as Setup 4-4, as the smallest single algorithm D-EER is twice as good.

B. Fusion-Groups of Size Fifteen

For fusion-groups of size fifteen random forest determines weights that lead to the same improvement as grid-search for fusion-groups of size four. However, in a significantly shorter period of time. Grid-search might lead to good results as well for fusion-groups of size fifteen. However, it would take more than 1100 years to run our approach for fusion-groups of size ten. Therefore, we were limited to normalising the scores and using weights determined by random forest for fusion-groups of size fifteen. As can be seen in Table IV the five most promising results all achieve an D-EER below 0.5% and all setups contain S-MAD and D-MAD scores. In our experiments the most promising results were achieved by not only fusion different feature extractors and classifiers but also S-MAD and D-MAD. However, a cause-effect relation cannot be made for certain as a significant deterioration in the D-EER could observed when fusing scores of S-MAD and D-MAD in setups using texture descriptors and landmark-based feature extractors. Since most of the scores used in the fusion-groups of size fifteen are part of every of those fusion-group, the results are similar.

VI. Conclusion

Our results suggest that fusing deep features and texture descriptors leads to the most promising results, while fusing only scores of the same feature extractor - generated using different classifiers and hyperparameter optimisation methods - mostly did not result in any improvement.

The results of the unweighted approach could not compete with those of the weighted approaches. For grid-search weights the most promising results were achieved fusing groups of four and for random forest when fusing groups of fifteen. Both achieve a relative improvement of -71.43% compared to the best score in the fusion-group. However, random forest is less time consuming than grid-search. In conclusion, if less scores are available, using grid-search to adjust the weights is feasible up to a group size of four and achieves good results, if, however, time is a relevant factor and more scores are available random forest can be used to achieve equally good results. The results of our weighted experiments indicate a significant improvement compared to individual MAD methods and compared to our unweighted results.

VII. Future Work

Due to time constraints we did not adjust random forest's scoring function, however, our results suggest that doing so could improve the random forest results. Hence, a time-efficient alternative to grid-search could be created. Furthermore, a validation of our results on a larger scale dataset would be interesting. In order to obtain such results a C++-implementation of our MAD method has been submitted to the NIST FRVT MORPH benchmark [25]. The MIPGAN-approach for generating morphed facial images introduced in [26] will be part of the NIST FRVT MORPH benchmark report. The results of the submitted method should therefore give an indication for its robustness towards GAN-generated morphs.

We only fused models based on features extracted by texture descriptors, deep features and landmarks. However, various other approaches exist and seem promising e.g. image noise pattern based feature extractors. Including these could positively influence the results and improve the D-EER even further.

References

[1] M. Ferrara, A. Franco, and D. Maltoni, "The magic passport," in *IEEE International Joint Conference on Biometrics*. IEEE, pp. 1–7. [Online]. Available: http://ieeexplore.ieee.org/document/6996240/

[2] U. Scherhag, C. Rathgeb, J. Merkle, R. Breithaupt, and C. Busch, "Face recognition systems under morphing attacks: A survey," vol. 7, pp. 23 012–23 026. [Online]. Available: https://ieeexplore.ieee.org/document/8642312/

[3] ISO/IEC 30107-3:2017, "Information technology — Biometric presentation attack detection — Part 3: Testing and reporting," International Organization for Standardisation, Tech. Rep., 2017.

[4] G. Heusch and S. Marcel, *Handbook of Biometric Anti-Spoofing*, 2nd ed., S. Marcel, M. S. Nixon, J. Fierrez, , and N. Evans, Eds., 2019.

[5] U. Scherhag, D. Budhrani, M. Gomez-Barrero, and C. Busch, "Detecting morphed face images using facial landmarks," in *Image and Signal Processing*, A. Mansouri, A. El Moataz, F. Nouboud, and D. Mammass, Eds. Springer International Publishing, 2018, vol. 10884, pp. 444–452. [Online]. Available: http://link.springer.com/10.1007/978-3-319-94211-7_48

[6] L. Debiasi, U. Scherhag, C. Rathgeb, A. Uhl, and C. Busch, "PRNU-based detection of morphed face images." IEEE, 2018, pp. 1–7. [Online]. Available: https://ieeexplore.ieee.org/document/8401555/

[7] U. J. Scherhag, "Morphing attacks and morphing attack detection," PhD dissertation, Technische Universität Darmstadt, 2018.

[8] U. Scherhag, L. Debiasi, C. Rathgeb, C. Busch, and A. Uhl, "Detection of face morphing attacks based on PRNU analysis," vol. 1, no. 4, pp. 302–317. [Online]. Available: https://ieeexplore.ieee.org/document/8846232/

[9] 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 *2017 International Conference of the Biometrics Special Interest Group (BIOSIG)*. IEEE, pp. 1–7. [Online]. Available: http://ieeexplore.ieee.org/document/8053499/

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

[11] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2015. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2015.7298682

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 2672–2680. [Online]. Available: http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf

[13] X. Yu, G. Yang, and J. Saniie, "Face morphing detection using generative adversarial networks," in *2019 IEEE International Conference on Electro Information Technology (EIT)*, 2019, pp. 288–291, ISSN: 2154-0373, 2154-0357.

[14] 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.

[15] U. Scherhag, C. Rathgeb, and C. Busch, "Morph deterction from single face image: a multi-algorithm fusion approach," in *Proceedings of the 2018 2nd International Conference on Biometric Engineering and Applications - ICBEA '18*. ACM Press, pp. 6–12. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3230820.3230822

[16] A. Makrushin, C. Kraetzer, J. Dittmann, C. Seibold, A. Hilsmann, and P. Eisert, "Dempster-shafer theory for fusing face morphing detectors," in *2019 27th European Signal Processing Conference (EUSIPCO)*. IEEE, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/8902533/

[17] S. Venkatesh, R. Ramachandra, K. Raja, and C. Busch, "Single image face morphing attack detection using ensemble of features," in *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*. IEEE, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/9190629/

[18] U. Scherhag, C. Rathgeb, J. Merkle, and C. Busch, "Deep face representations for differential morphing attack detection." [Online]. Available: http://arxiv.org/abs/2001.01202

[19] N. Damer, A. Opel, and A. Nouak, "Biometric source weighting in multi-biometric fusion: Towards a generalized and robust solution," in *2014 22nd European Signal Processing Conference (EUSIPCO)*, 2014, pp. 1382–1386.

[20] N. Damer, S. Zienert, Y. Wainakh, A. Moseguí Saladié, F. Kirchbuchner, and A. Kuijper, "A Multi-detector Solution Towards an Accurate and Generalized Detection of Face Morphing Attacks."

[21] 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, 1998.

[22] 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 *Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[23] ISO/IEC JTC1 SC37 Biometrics, "International standards ISO/IEC TR 24722, multimodal and other multibiometric fusion," International Organization for Standardisation, Tech. Rep., 2015.

[24] M. Singh, R. Singh, and A. Ross, "A comprehensive overview of biometric fusion," *Information Fusion*, vol. 52, pp. 187 – 205, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S156625351830839X

[25] M. Ngan, P. Grother, K. Hanaoka, and J. Kuo, "Face recognition vendor test (FRVT) part 4:: MORPH - performance of automated face morph detection," p. NIST IR 8292. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/ir/2020/NIST.IR.8292.pdf

[26] H. Zhang, S. Venkatesh, R. Ramachandra, K. Raja, N. Damer, and C. Busch, "Mipgan – generating strong and high quality morphing attacks using identity prior driven gan," 2021.