Face Morphing Attack Detection

Szymon Żmijewski[1]

Abstract: As tools designed for face image manipulation are getting popular, the need for robust detection software is becoming more urgent. One of the main threats is face morphing. As new morphing attack detection (MAD) methods appear, the problem of reliable and standardized benchmarking and testing remains rarely tackled. This paper aims to implement one of the state-of-the-art approaches in order to discover its characteristics and provide more data regarding its robustness. The explored method uses Deep Face Representation retrieved from publicly available tool ArcFace, which is used for feature combination, Support Vector Machine model designing and classification. The model uses FERET database for training and FRGCv2 database for testing. As a result a promising Detection Equal Error Rate around 3% is reported and DET curves are provided.

Keywords: Differential Face Morphing Attack Detection, D-MAD, Deep face representation, Support Vector Machine,

[1] Technical University of Denmark, DTU Compute, Copenhagen, Denmark, s222919@dtu.dk

1 Introduction

The threat of morphing attacks on face recognition systems, including Automatic Border Control (ABC) gates, is a well known issue. It has been proven that detecting such a morphed picture can be difficult for the human eye. Thus a whole research area regarding automatic morph attach detection (MAD) came to existence. A number of methods were proposed, introducing two main approaches: namely Single Image Based MAD (S-MAD), trying to determine if a probe picture is a morph based exclusively on information contained in the given image, and Differential MAD (D-MAD), using a reference picture (e.g. captured on the ABC gate).

However, both approaches suffer from the lack of relevant training data to prove their robustness. Thus, they tend to have a limited scope or strict requirements as to how the pictures should be obtained or processed. In a real-life scenario some pose variations, different lighting, varying facial expressions etc. must be assumed. Some models suggested in the literature may fail to meet the expectations when tested with a dataset different from the original. An important role in that topic is played by the software that is used to create a morph. Several tools became especially popular: FaceFusion[2], FaceMorpher[3], OpenCV[4] and UBO-Morpher [5] (ubo).

In spite of the existing benchmarking standards, a variety of ways to present the effectiveness of methods is present across different research papers. The purpose of this research is to reproduce one of the state-of-the-art methods, verify its effectiveness and report the results according to ISO/IEC 30107-3 standards containing:

1. Detection Error Tradeoff (DET) curves [Ma97].
2. Detection Equal Error Rate (D-EER): the operating point in which APCER and BPCER are equal.

This research focuses on the evaluation of D-MAD methods since they have proven to often perform better [RB] because of the additional information in the trusted reference picture obtained for example in the ABC gate scenario.

The rest of this paper is organized as follows. Section 2 introduces the variety of methods together with the method chosen for reproduction. Then in section 3 the reproduction and verification process is described. Subsequently, section 4 presents obtained results. Section 5 discusses the outcome of performed research and compares the results with ones obtained using other methods. Finally, section 6 draws a conclusion and summarizes the work.

[2] http://www.wearemoment.com/FaceFusion/
[3] github.com/alyssaq/face_morpher
[4] www.learnopencv.com/face-morph-using-opencv-cpp-python/
[5] http://biolab.csr.unibo.it/research.asp?organize=Activities&select=&selObj=220

2 Background

S. Venkatesh et al. [RB] provided a broad overview of available methods for both S-MAD and D-MAD. They empathize the importance of comprehensive testing and benchmarking of all algorithms. They also express the need for a standardized approach to reporting the results. The aim of this paper is to further analyze one of the methods and show its efficiency in a standardized manner.

In general, D-MAD algorithms could be divided into 2 main categories:

1. Feature-based - in which a feature vector is extracted from both probe and *bona fide* images and the decision is based on detected differences between the two
2. Demorphing - by trying to revert the process of face morphing to obtain an image that does not contain features added by the individual's face. Then a classic facial verification algorithm should report a significant dissimilarity of the *bona fide* image and the reference image

Comparing features of images is constrained by the type of image and the segmentation of a face on an image, but it performs well on varying image resolutions and qualities.

The latter carries a number of important limitations, namely more constraints to the image format, higher sensitivity to pose and lighting changes and a known a priori blending factor.

Although feature-based approach is not flawless, its restrictions seem more feasible to overcome. For this reason, the research will be carried out on one of the methods from this category.

2.1 State-of-the-art methods

Feature-based morphing detection algorithms currently mostly make a use of neural networks or machine learning. Some of them utilize landmark detection [Da19] [RRB16], other extract features using texture descriptors [SRB18] and a group of methods involves deep learning [Se20] [Si19]. Finally, another approach is to retrieve features from deep face representation which was explored by Scherhag et. al [Sh22].

2.2 Reproduced method

The last method will be a subject of verification and testing in the experiment. The main idea behind it is to use a tool designed for automatic facial recognition as a feature extractor and by calculating the difference between two vectors (assuming that one of them is retrieved from a trusted source) create a value will be then used as an input for a machine learning algorithm whose task is to classify it as either morphed image or a *bona fide*. This definition of the method still enables a lot of flexibility and the chosen approach will be described in the following sections.

2.3 Research questions

Based on the challenges introduced so far, the following research questions are formulated.

RQ1 Can the way of calculating the distance between the feature vectors impact the results of the model quality?

RQ2 Are any of the morphing tools sufficient to train the model for recognizing morphs regardless of used face morphing software?

RQ3 Is the model better when trained on a collection of images produced using all the tools or just one of them?

RQ1 is supposed to verify if different dissimilarity scores between a given pair of feature vectors can significantly influence how accurate the classifier becomes. **RQ2** touches upon the mentioned problem of the lack of training data. Proving that one of the tools can be used as a generalization of the others can simplify the efforts of training D-MAD models and allow to trust the obtained results more. **RQ3** is expanding on **RQ2** by checking if training the model on just one morphing tool is actually better than using all of them. This idea comes from the intuition that it should be easier to distinguish a morph from a *bona fide* sample when using only one tool since the morphs should have more common features.

3 Methods and Materials

The conducted research consisted of several stages. First, the dataset was analyzed and explored with special attention to how the images containing morphed faces are organized. Then the pipeline introduced in the original paper was reproduced, already considering the conclusion from the initial approach. Subsequently, the experiments necessary to answer the Research Questions were run to then benchmark implemented techniques against each other and the results obtained in the literature.

3.1 Dataset

Two commonly used databases were used as a subject of the experiment: FRGCv2 [Ph05] and FERET [Ph00]. They both contain images produced using the aforementioned face morphing software and include over 2400 probe images in total (while more than one image could be obtained from a single individual).

In order to unify the experiments and enable a comparison with the original method, the same division for training and testing data was implemented (using the whole FRGCv2 database as the training set and the whole FERET database as the testing set). This ensures a strict separation of the datasets, avoiding over-fitting, while increasing the chances of an error-less classification of the images (as seen in Section 3.2).

The example morphs are represented in Figure 1 and 2.

Fig. 1: An example morphed face (middle) produced from two *bona fide* samples (left and right) from FERET database using Facefusion tool.

Fig. 2: An example morphed face (middle) produced from two *bona fide* samples (left and right) from FRGC database using Facefusion tool.

3.2 Reproduction

Based on the results obtained by Scherhag et al., it was decided to focus on the most promising tools and ideas. Thus, only one facial recognition system was involved: ArcFace [De19] which provides a deep face representation as a vector containing 512 features. At this stage, an observation made in the original paper that the difference vectors are indeed sufficient to classify probe images was confirmed. This step of the research was repeated in order to develop on top of it. The results are represented in Figure 3.

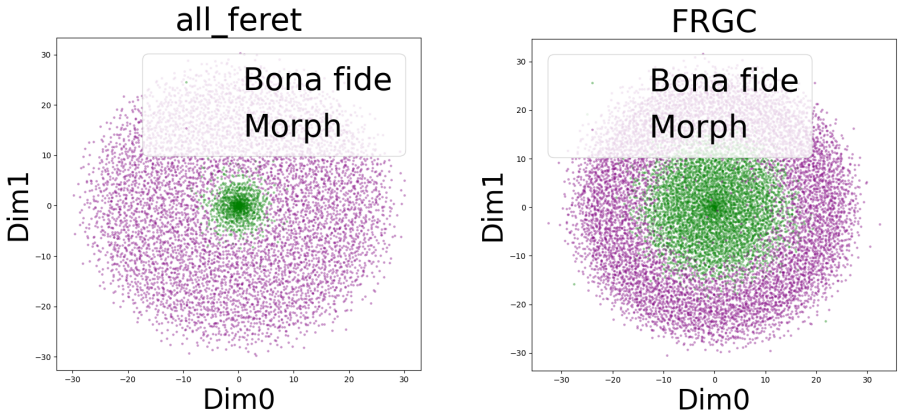

Fig. 3: Difference vectors obtained from FERET (left) and FRGCv2 (right) after applying a Multi-dimensional Scaling (MDS), reducing the number of dimensions of the vectors from 512 to 2.

The process of visualizing the difference vectors was performed on images produced using individual tools (contrary to the original set where the whole database was taken into account). These difference vectors are visualized in the same fashion on Figure 4. It is noticeable how an error-less separation is easier when only one morphing algorithm is used at a time. It is worth noting that images from one database are considered (so similarities in face poses and lighting can be assumed), making the classification easier and potentially improving the ML model.

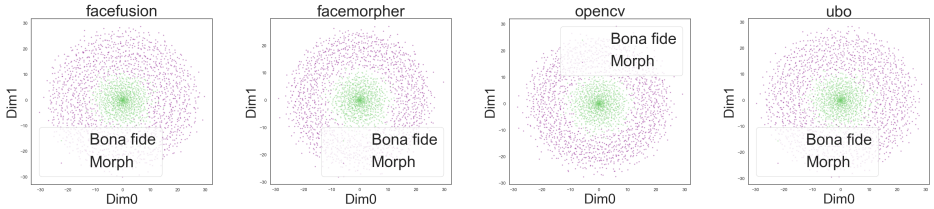

Fig. 4: Difference vectors obtained from FERET after applying a Multidimensional Scaling (MDS), reducing the number of dimensions of the vectors from 512 to 2 for individual face morphing tools accordingly.

Again, drawing from the initial research, only SVM was used as a classifier for the tests, since it has provided the best results.

3.3 Experiments

To answer **RQ1**, it was decided to calculate the similarity between feature vectors as cosine distance. The reason for this metric is that the comparison score is expressed as only one number (instead of a difference vector that preserves the dimensionality of the original feature vector) while considering as many of the vector's characteristics, as possible. Although it does not cover all possible ways of retrieving a distance between two vectors, it will show how different the results can be because of using different techniques.

A visualization of a difference between feature vectors expressed as cosine distance was performed and represented in Figure 5 and 6.

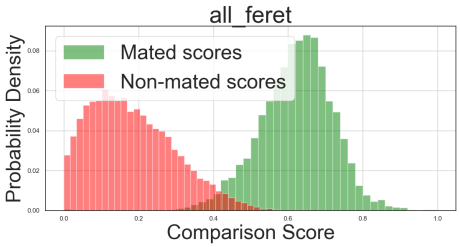

Fig. 5: Difference between vectors obtained from the whole FERET database expressed as cosine distance.

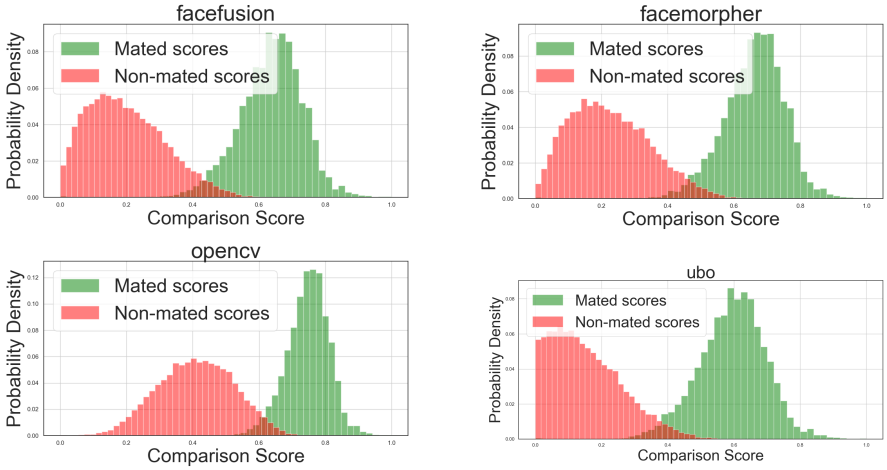

Fig. 6: Difference between vectors obtained from FERET database expressed as cosine distance for individual face morphing tools accordingly.

That proves that it is reasonable to assume that a simple cosine distance can be sufficient for the classification process. Thus, both distance methods will be subject to further testing.

For both of the distance metrics and for each training set (for the whole FERET database and for each face morphing tool) a set of hyper-parameters for SVM was tested. The values providing the results with the highest accuracy are presented in Table 1 and 2.

hyper-parameter	all	facefusion	facemorpher	opencv	ubo
C	1	1	0.1	10	1
gamma	0.001	0.001	0.001	0.001	0.001

Tab. 1: Hyper-parameters providing the highest accuracy for difference vector-based SVM.

hyper-parameter	all	facefusion	facemorpher	opencv	ubo
C	1	1	0.1	0.1	100
gamma	1	1	1	0.1	1

Tab. 2: Hyper-parameters providing the highest accuracy for cosine distance-based SVM.

Since the difference vectors have 512 dimensions, it is not feasible to represent False Match Rate (FMR) and False Non-Match Rate (FNMR). To tackle this, for the trained model distances between the difference vectors from the test dataset and the model's hyperplane were calculated. The results were normalized and projected onto a histogram shown on Figure 7.

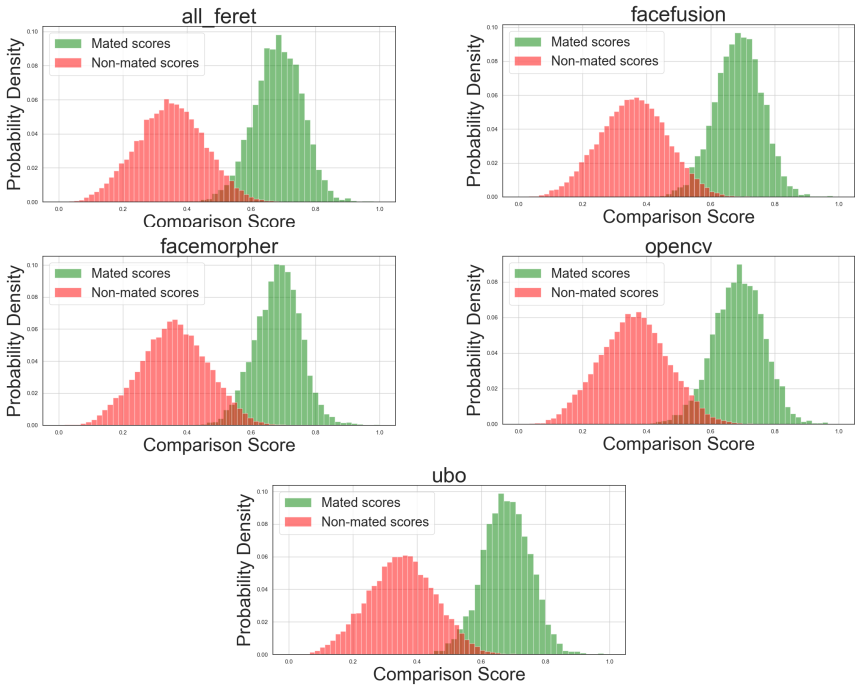

Fig. 7: Comparison scores retrieved from difference vectors obtained from the FERET database (combined and for each of the morphing tools) expressed as the distance from the SVM's hyperplane.

4 Results

All of the models were tested and their quality was assessed in terms of prediction accuracy. The obtained values are shown in Table 3.

	all	facefusion	facemorpher	opencv	ubo
difference vector	0.9034	0.9247	0.9337	0.9454	0.9231
cosine distance	0.8892	0.9136	0.9467	0.9447	0.9153

Tab. 3: Hyper-parameters providing the highest accuracy for cosine distance-based SVM.

For both dissimilarity metrics, the Detection Error Tradeoff (DET) [Ma97] curves are depicted in Figure 8.

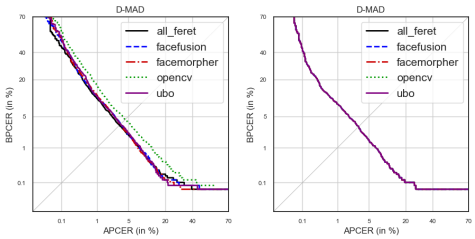

Fig. 8: DET curves for difference vector dissimilarity metric (left) and the cosine distance dissimilarity metric (right).

It is worth noticing that all curves for cosine distance case are almost identical. This was verified and confirmed, that the values for all of the dissimilarity scores for this scenario were substantially shifted.

The error rates (D-EER) for the classifiers compared to the ones obtained in the original paper using the same tool and datasets in order to express the flexibility of both when paired with different feature processing approaches. are summarised in Table 4.

MAD	D-EER(%)
proposed model with difference vectors trained on whole FERET	3.32
proposed model with difference vectors trained on facefusion	3.54
proposed model with difference vectors trained on facemorpher	3.38
proposed model with difference vectors trained on opencv	4.30
proposed model with difference vectors trained on ubo	3.48
proposed model with cosine distances trained on whole FERET	3.09
proposed model with cosine distances trained on facefusion	3.09
proposed model with cosine distances trained on facemorpher	3.09
proposed model with cosine distances trained on opencv	3.09
proposed model with cosine distances trained on ubo	3.09
MFN [Da22]	16.36
XN [Ch17]	9.75
HRN [Wa20]	10.89

Tab. 4: Hyper-parameters providing the highest accuracy for cosine distance-based SVM.

5 Discussion

In the context of the study, we formulated the following research questions:

RQ1 Can the way of calculating the distance between the feature vectors impact the results of the model quality?
Yes, even defining the software used to extract the features of an image and restricting feature processing to calculating a distance between two vectors, there is still enough flexibility left to significantly influence the model's performance in terms of morph detection accuracy. However, the complexity of a metric is not as relevant as it may initially seem. Using the cosine distance with just 1 dimension performed just as well or better than the difference vector with 512 dimensions both in terms of the prediction accuracy (avg. 0.922 and 0.926 accordingly) and D-EER (avg. 3.09 and 3.60 accordingly).

RQ2 Are any of the morphing tools sufficient to train the model for recognizing morphs regardless of used face morphing software?
Yes, analyzing the DET curves a conclusion can be drawn that testing the model on images morphed using OpenCV results in a slightly worse performance. All of the other morphing tools reported similar results and can be considered sufficient for training a model designed to later classify images with morphed faces generated using all of the mentioned morphing tools. This finding may be a step towards more reliable D-MAD testing and benchmarking since it proves that the concerns about the lack of a unified approach to this process may not be that restricting.

RQ3 Is the model better when trained on a collection of images produced using all the tools or just one of them?
Drawing from the conclusions to RQ2, the quality can be treated as equal for all approaches except for a model trained on OpenCV, which is a significant finding since it suggests that results obtained by using different techniques can be directly compared to each other.

The quality of the represented pipeline is not only efficient but also easy to implement by using publicly available solutions and libraries, while maintaining reasonably low computational complexity of feature extraction and ML model training.

It is also important that for each model explored in the research, both the training data and hyper-parameters varied. The reported quality applies only to individually treated SVMs. Thus, while using a common model for all training datasets, a quality trade-off should be expected.

6 Conclusions

All posed research questions were answered and the results prove to be promising. They show that the suggested direction is worth exploring.

The original method was successfully reproduced and the intended changes were introduced showing a possible distinction between morphed faces and *bona fide* images in Section 3.3. The process of hyper-parameter tuning for SVM model was described, enabling the reproduction of the conducted experiments. Then the results were reported in accordance to ISO/IEC 30107-3 standards in Section 4.

There is room for further research in terms of finding the optimal vector distance metric and improving the introduced model. Potentially another face recognition software can be utilized to extract face feature vectors which, in combination with either already explored or a new distance metric can provide better quality results.

It is also still viable to test the approach using other realistic databases, however, the strict separation between the training and testing data (in terms of using two different databases with different capturing devices and characteristics of the images) ensured that any overfitting was avoided, making the model more robust and the results more trustworthy.

References

[Ch17] Chollet, François: , Xception: Deep Learning with Depthwise Separable Convolutions, 2017.

[Da19] Damer, Naser; Boller, Viola; Wainakh, Yaza; Boutros, Fadi; Terhörst, Philipp; Braun, Andreas; Kuijper, Arjan: Detecting Face Morphing Attacks by Analyzing the Directed Distances of Facial Landmarks Shifts. In (Brox, Thomas; Bruhn, Andrés; Fritz, Mario, eds): Pattern Recognition. Springer International Publishing, Cham, pp. 518–534, 2019.

[Da22] Damer, Naser; López, César Augusto Fontanillo; Fang, Meiling; Spiller, Noémie; Pham, Minh Vu; Boutros, Fadi: , Privacy-friendly Synthetic Data for the Development of Face Morphing Attack Detectors, 2022.

[De19] Deng, Jiankang; Guo, Jia; Xue, Niannan; Zafeiriou, Stefanos: ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4685–4694, 2019.

[Ma97] Martin, A.; Doddington, G.; Kamm, Terri; Ordowski, M.; Przybocki, Mark: The det curve in assessment of detection task performance. The DET Curve in Assessment of Detection Task Performance, pp. 1895–1898, 01 1997.

[Ph00] Phillips, P.J.; Moon, Hyeonjoon; Rizvi, S.A.; Rauss, P.J.: The FERET evaluation methodology for face-recognition algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(10):1090–1104, 2000.

[Ph05] Phillips, P.J.; Flynn, P.J.; Scruggs, T.; Bowyer, K.W.; Chang, Jin; Hoffman, K.; Marques, J.; Min, Jaesik; Worek, W.: Overview of the face recognition grand challenge. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). volume 1, pp. 947–954 vol. 1, 2005.

[RB] Raja, Sushma Venkatesh Raghavendra Ramachandra Kiran; Busch, Christoph: Face Morphing Attack Generation & Detection: A Comprehensive Survey.

[RRB16] Ramachandra, Raghavendra; Raja, Kiran; Busch, Christoph: Detecting Morphed Face Images. 09 2016.

[Se20] Seibold, Clemens; Samek, Wojciech; Hilsmann, Anna; Eisert, Peter: Accurate and robust neural networks for face morphing attack detection. Journal of Information Security and Applications, 53:102526, 2020.

[Sh22] Shiqerukaj, E.; Rathgeb, C.; Merkle, J.; Drozdowski, P.; Tams, B.: Fusion of Face Demorphing and Deep Face Representations for Differential Morphing Attack Detection. 2022. Cited by: 0.

[Si19] Singh, Jag Mohan; Ramachandra, Raghavendra; Raja, Kiran B.; Busch, Christoph: Robust Morph-Detection at Automated Border Control Gate Using Deep Decomposed 3D Shape Diffuse Reflectance. In: 2019 15th International Conference on Signal-Image Technology Internet-Based Systems (SITIS). pp. 106–112, 2019.

[SRB18] Scherhag, Ulrich; Rathgeb, Christian; Busch, Christoph: Towards Detection of Morphed Face Images in Electronic Travel Documents. In: 2018 13th IAPR International Workshop on Document Analysis Systems (DAS). pp. 187–192, 2018.

[Wa20] Wang, Jingdong; Sun, Ke; Cheng, Tianheng; Jiang, Borui; Deng, Chaorui; Zhao, Yang; Liu, Dong; Mu, Yadong; Tan, Mingkui; Wang, Xinggang; Liu, Wenyu; Xiao, Bin: , Deep High-Resolution Representation Learning for Visual Recognition, 2020.