Impact and Mitigation of Quality Degradation for Differential Morphing Attack Detection

Torsten Schlett, Christian Rathgeb, Juan E. Tapia and Christoph Busch
da/sec - Biometrics and Security Research Group
Hochschule Darmstadt, Germany / European University of Technology, European Union
Email: {torsten.schlett, christian.rathgeb, juan.tapia-farias, christoph.busch}@h-da.de

***Abstract*—Differential Morphing Attack Detection (D-MAD) aims to detect morphed face images by comparing the suspected potential morph against a trusted live capture. While the quality of a suspected image used in a passport or similar will typically be controlled to avoid degradation by environmental factors such as lighting, the quality of the trusted live capture could more easily vary in real conditions. This paper examines how a D-MAD system is affected by varying trusted live capture quality, how the same quality variations impact Face Image Quality Assessment (FIQA), and demonstrates how said FIQA can be utilized to substantially mitigate the D-MAD performance impact via another model. The experiments in particular consider four synthetic and thus clearly controlled image defect types, two corresponding to environmental lighting variation and two to blur, all based on approaches from the NIST FATE Quality SIDD report. The tested D-MAD system is based on deep face representations and the tested FIQA algorithms are parts of the recently established OFIQ project.**

***Index Terms*—Biometrics, face recognition, face image quality assessment, differential morphing attack detection, defects.**

I. Introduction

Morphing attacks on face recognition [1] systems first combine data from multiple biometric subjects into one morphed face image, which is then enrolled in the system, e.g. as part of a passport application. In practice the face recognition system will compare the enrolled data against a trusted live capture, i.e. a face image that cannot be another morphed image. A Differential Morphing Attack Detection (D-MAD) system likewise compares these image pairs, except to predict whether the suspected image used for enrolment is a morph or bona fide (not morph) image. The quality of a face image may however vary due to various environmental factors such as lighting, which can be analysed by existing Face Image Quality Assessment (FIQA) algorithms [2]. Usually quality degradations are more likely for the trusted live capture images taken e.g. at an automated border control gate, whereas the enrolment process can expect and check for higher image quality. The contributions of this work can be summarized as follows:

- The impact of synthetic environmental degradation (subsection III-B) of trusted live capture face images (subsection III-A) on the performance of a D-MAD system (subsection III-C) is evaluated (subsection IV-B).
- The response of a set of Face Image Quality Assessment (FIQA) algorithms (subsection III-D) to the same synthetic face image degradation is tested as well (subsection IV-C).
- Based on the calculated potential for D-MAD decision threshold optimization (subsection IV-B) and the tested FIQA algorithms (subsection IV-C), a lightweight D-MAD threshold optimization model is developed to noticeably mitigate the face image degradation impact on the D-MAD performance (section V). See Figure 1.

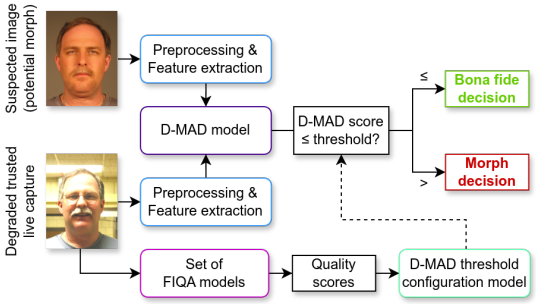

Fig. 1. An overview of the D-MAD process with a threshold optimization model that is using a set of quality scores for the trusted live capture as input.

II. Related work

Our work is closely related and complementary to the recent work by Franco et al. [3], which likewise investigated the impact of face image quality variations on D-MAD. In the context of automated border control, the analysis in [3] concluded that D-MAD performance mostly depends on the quality of the "gate" image (which corresponds to the "trusted live capture" in our work), as opposed to the quality of the "enrolment" image (the "suspected" image in our work). The differences between [3] and our work include the used evaluation data, the selection of algorithms, and the analysis approach. While [3] investigated the impact of illumination uniformity, focus (i.e. sharpness), yaw, pitch and roll defects, our work is focused on Gaussian blur, motion blur, overexposure and underexposure as defects of interest. More specifically regarding the quality impact analysis, our work employs synthetic quality degradation across different levels of severity to investigate a selection of defect types in a controlled manner (subsection III-B), whereas [3] investigated a broader set of FIQA and D-MAD algorithms on data without strictly controlled quality degradation levels for the individual defect types prior to FIQA (not to be confused with [3] using quality assessment output to group images into quartiles for the analysis). In this regard our work supports [3] by confirming that quality degradation can substantially affect D-MAD performance. But beyond providing a different kind of quality impact analysis (section IV), our work also investigates the use of FIQA to mitigate the D-MAD performance degradation via automatic D-MAD decision threshold optimization, with promising results for the examined setup (section V).

TABLE I
THE NUMBER OF SUBJECTS, IMAGES, AND D-MAD PAIRS IN THE USED DATASET (SUBSECTION III-A).

Dataset	Subjects	Images	Pairs
FERET subset	529	3431	7066
FRGCv2 subset	533	6562	29230
Combined	1062	9993	36296

TABLE II
THE NUMBER OF TRUSTED LIVE CAPTURE IMAGES, THE NUMBER OF SUSPECTED BONA FIDE OR MORPH IMAGES (FOR EACH MORPH TYPE), AND THE NUMBER OF BONA FIDE OR MORPH D-MAD IMAGE INPUT PAIRS IN THE USED DATASET (SUBSECTION III-A).

	Images			Pairs	
		Suspected			
Dataset	Trusted	Bona fide	Morph	Bona fide	Morph
FERET subset	786	529	529	786	1570
FRGCv2 subset	1724	982	964	3294	6484
Combined	2510	1511	1493	4080	8054

Two earlier and comparatively somewhat less closely related works by Fu et al. [4] [5] investigated the separability of (F)IQA output for bona fide vs. morphed images, including the possibility to repurpose (F)IQA approaches for MAD based on that separability. One of the more promising results were obtained by the MagFace [6] approach, which is considered in our work as well (see subsection III-D, but note that this may not be the exact same MagFace model).

Another more indirectly related work by Borghi et al. [7] introduced "Video-based Morphing Attack Detection (V-MAD)", which considers multiple video frames as the trusted live capture input for D-MAD. The final V-MAD score output is then produced by one of various examined fusion strategies. A subset of these fusion strategies employed FIQA, either by selecting the best image among the given video frames, by forming a weighted average of the D-MAD scores across the frames, or by feeding both the frames' D-MAD scores and the quality scores to another machine learning model. While our work investigates typical D-MAD instead of V-MAD, it is related to [7] insofar that we likewise research the use of FIQA to enhance (D-)MAD performance via decision threshold optimization (section V), which could potentially also be used to design another V-MAD fusion strategy in future work.

III. EXPERIMENT SETUP

A. Dataset

The experiments in this paper use the MAD dataset created by Scherhag et al. in [9], more specifically the variant without image post-processing. This dataset is based on bona fide (i.e. non-morph) face images from FERET [10] [11] and FRGCv2 [12]. Pairs of bona fide images were used to generate morphed images with four different morphing algorithms, namely the proprietary "FaceFusion", the open source "FaceMorpher", a self-made "OpenCV"-based approach, and "UBO-Morpher" from the University of Bologna [9] [13]. For further details and morphing algorithm example images see [9].

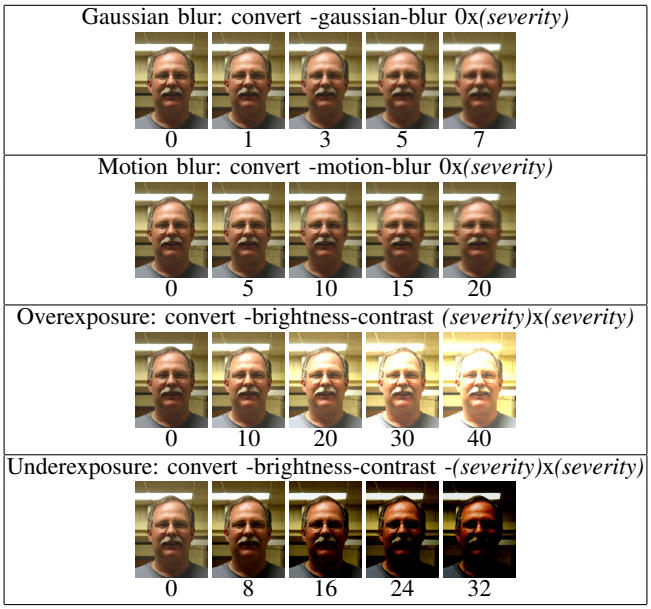

Fig. 2. Example images for the four synthetic defect types, with all used degradation severity steps (0 being no degradation, i.e. the original image independent of the defect type). The ImageMagick [8] "convert" commands are shown after the title of each type.

The D-MAD experiments require image pairs, with one image representing the trusted live capture image (i.e. an image known to be bona fide, which may vary in quality), and the other image representing the suspected image which is either bona fide or morphed (of presumably good quality). The set of trusted live capture images and the set of suspected images are disjoint. Table I shows the number of subjects, images, D-MAD pairs, and Table II provides additional details regarding the number of trusted/suspected bona fide/morphed images and the bona fide/morphed D-MAD pairs. The width and height of all images is 720 and 960, approximately.

B. Synthetic degradation

Synthetic degradation for four defect types is utilized to approximate potential environmental degradations of the trusted live capture images in a real scenario, such as automated border control. The concrete approaches are based on configurations described in NIST FATE Quality SIDD report version 2024-04-26 [14], which used them for defect evaluations: "Gaussian blur" ("Resolution" in [14]) to represent e.g. camera defocus, "Motion blur", "Overexposure", and "Underexposure". Simple ImageMagick [8] "convert" commands are used for all four defect types. The command for each synthetic defect type is parameterised by a single "degradation severity" integer, for which higher values correspond to worse image quality.

For the sake of computational efficiency four severity steps are used for each defect type, and these steps are again selected within the corresponding ranges used by NIST in [14]. Figure 2 shows these used degradation severity steps with example images, alongside the original image ("0" degradation severity).

Note that the same severity steps for different defect types do of course not necessarily correspond to the same face image quality degradation according to various FIQA algorithms, hence the FIQA impact evaluation in subsection IV-C. Nevertheless, the severity steps for the two blur defect types appear to result in approximately comparable blur strength, as indicated by the example images in Figure 2.

C. D-MAD system

The used D-MAD system is a newer variant of the approach introduced by Scherhag et al. [9]. It corresponds to the NIST FATE MORPH [15] submission with the identifier "hdadfr-006". This type of D-MAD system primarily involves a model trained to asymmetrically compare face recognition feature vectors from a trusted live capture and a suspected image. Further details will be described as part of the evaluation subsection IV-B.

D. Quality assessment

The used Face Image Quality Assessment (FIQA) algorithms are measures provided by the "Open Source Face Image Quality" (OFIQ) project [16] [17]. The concrete version used in this paper is the GitHub repository 2024-05-10 state the OFIQ Python adapter fork [18]. More specifically, the following OFIQ measures, which presumably are more relevant to the considered defect types (subsection III-B), are utilized:

- Sharpness: For the synthetic defect types "Gaussian blur" and "Motion blur". This measure consists out of a hand-crafted part followed by a random forest model.
- Over-Exposure-Prevention: For the defect type "Overexposure". This is a hand-crafted measure.
- Under-Exposure-Prevention: For the defect type "Underexposure". This is a hand-crafted measure.
- Unified: For all defect types. This measure is a MagFace [6] model, i.e. a model simultaneously trained both for face recognition and for unified FIQA, with only the latter being relevant here.

OFIQ provides quality score output in the $[0, 100]$ integer range for all of these measures, so that higher values are intended to indicate better facial biometrics utility.

IV. Evaluation

A. Image processing failures

Both the D-MAD system and OFIQ can fail to process an image, and such failure cases did occur for some of the degraded images: For the defect type "Overexposure" there were approximately 0.01% OFIQ failure cases at degradation severity 40 (the highest used setting). For "Gaussian blur" there were approximately 0.05% and 0.27% D-MAD failure cases at severity 5 and 7, respectively. Lastly and most notably, for "Underexposure" comparatively larger percentages of failure cases occurred for both OFIQ and D-MAD, as shown in Figure 3. Images for which such complete processing failure cases occurred are simply not considered as part of the corresponding distributions in the following evaluation. The reasoning is that an automated system in an operational scenario should analogously deny the use of images that cannot be processed properly, instead of e.g. simply assuming worst-case values for such cases (especially for the D-MAD part). The overall conclusions that can be drawn from the following evaluations would however likely not change regardless, since arguably only the "Underexposure" defect type involves substantial failure case percentages. Besides complete image processing failure cases, individual OFIQ [16] measures may fail as well. For these cases the overall OFIQ image processing was technically successful and other relevant measures for the same image may not have failed, so the affected measures' output is set to the worst quality score, which is 0. In an operational scenario one image could then for example be discarded based on the quality scores from multiple OFIQ measures. Or these quality scores could be used to adjust the D-MAD score threshold. The latter approach is examined in section V.

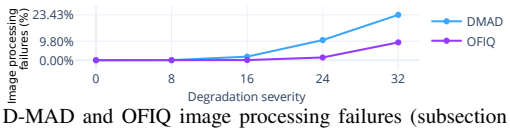

Fig. 3. D-MAD and OFIQ image processing failures (subsection IV-A) for the "Underexposure" defect type.

B. D-MAD impact

This part of the evaluation examines the impact of image degradation on the D-MAD decisions. D-MAD scores are floating-point values in the range $[0, 1]$, with higher values indicating a morphed image. These scores are computed by the D-MAD system previously described in subsection III-C, which compares potentially degraded but trusted live capture images against a corresponding suspected image of presumably good quality (e.g. a potentially morphed passport image). Binary D-MAD decisions are then derived from the D-MAD scores via a simple threshold comparison. The used default threshold for the examined system is 0.5, so that scores above that threshold imply a morphed image, while scores below or equal to that threshold conversely imply a bona fide image.

The evaluation results in Figure 4 show that image degradation of the trusted live capture tends to decrease the percentage of correct D-MAD decisions for bona fide images, while the percentage for morphed images increases. This is because the D-MAD decisions' underlying D-MAD scores tend to increase due to the image degradation. Of the defect types, "Underexposure" distinctly yielded the strongest impact. And "Gaussian blur" notably had much lower impact than "Motion blur" for higher severity settings, despite the seeming similarity of the blur strengths visible for example in Figure 2.

Note that the model which produces the D-MAD scores was trained in part on the evaluation dataset without degradation (subsection III-A), which means that it may be more effective here than it might be for other data. The evaluation results demonstrate that this potential advantage does not prevent performance impacts due to image degradation for this particular kind of D-MAD system. If you are interested in a general D-MAD performance evaluation of various systems, refer to NIST FATE MORPH [15] (the submission for the used D-MAD system is "hdadfr-006").

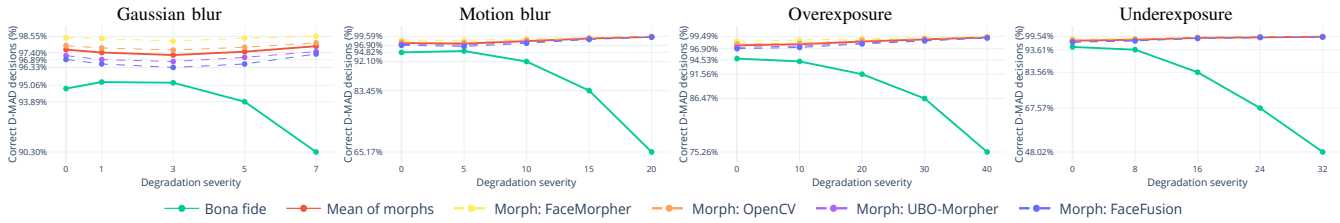

Fig. 4. Each plot shows the trusted live capture image degradation impact on the percentage of correct D-MAD decisions for one of the four defect types (see the title above each plot), using the default D-MAD decision threshold constant 0.5 described in subsection IV-B. The percentages are plotted on the Y-axis (higher percentages being better), while the degradation severity settings are plotted on the X-axis (increasing from left to right, 0 being no degradation).

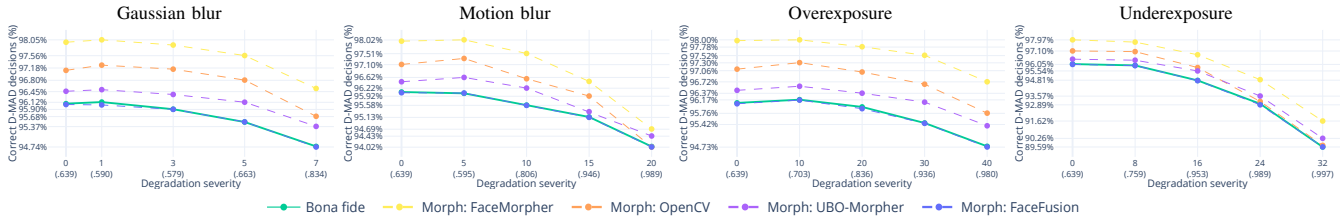

Fig. 5. Plots analogous to Figure 4, except that the D-MAD score thresholds used to form the binary D-MAD decisions (bona fide or morph) are optimized for each degradation severity per defect type via prior knowledge of the correct decisions, as described in subsection IV-B (the model-based optimization follows in Figure 7). The optimized threshold values are approximately indicated in brackets below the degradation severity labels on the X-axis.

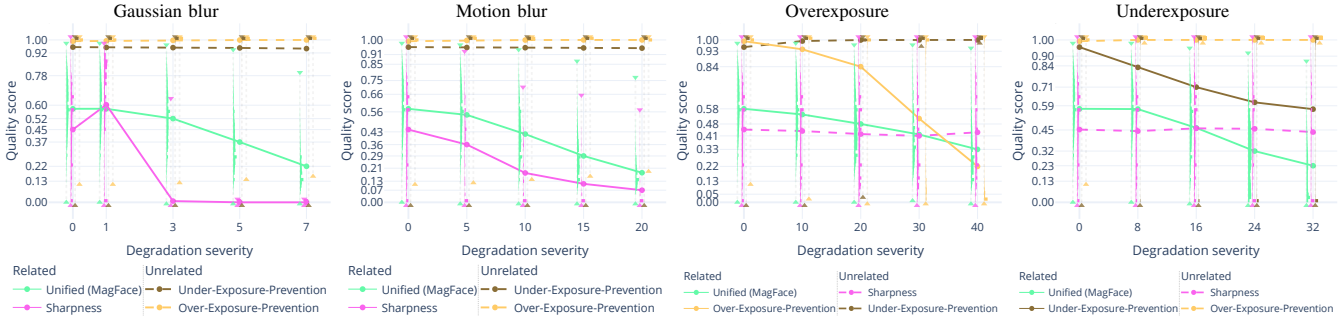

Fig. 6. Each plot shows the trusted live capture image degradation impact on the quality score output of the examined FIQA algorithms from subsection III-D (higher Y-axis values indicate better image quality) for one of the four defect types (see the title above each plot). As indicated in the legends below the plots, the FIQA algorithms are categorized as either "related" (solid curves) or "unrelated" (dashed curves) with respect to the defect type. "Related" FIQA algorithms are expected to respond to the degradation, while the "unrelated" ones are expected to be mostly unaffected. The curves show the mean quality score, triangles show the minima/maxima, and rotated (vertical) histograms additionally indicate the quality score distributions.

If the D-MAD threshold could be optimized separately for each trusted live capture, then the image degradation impact could be mitigated. This mitigation could be quite substantial, as Figure 5 illustrates in contrast to Figure 4. Figure 5 more specifically shows the D-MAD decision impact with thresholds optimized for each defect type severity, so that the worst correct decision percentage (i.e. the lowest Y-axis value among the curves) remains as high as possible. In other words, this figure shows the results using a different threshold for each degradation configuration to keep the worst-case as benign as possible, if the correct D-MAD decisions are already known. This knowledge is of course not available in an operational scenario, else the D-MAD system would not be required, but this data reveals to what extent such operational optimization may be possible. D-MAD threshold optimization that can actually be used in an operational scenario would have to e.g. base the optimization on quality assessment data for the trusted live capture, which is investigated in section V, using the thresholds computed here as model output training targets.

C. FIQA impact

Figure 6 shows the impact of the image degradation on the quality score output of the OFIQ [16] measure selected in subsection III-D, whereby the curves are grouped by the measures being either "related" or "unrelated" to each defect type. This analysis is analogous to the D-MAD impact analysis from the prior subsection IV-B in terms of the defect types and severity settings, but here the Y-axis represents the quality scores scaled to the $[0, 1]$ range, with lower quality indicating stronger degradation.

If the impact on the quality scores correlates with the impact on the D-MAD scores across the degradation configurations, then it should be possible to use these quality scores to e.g. only allow images with a certain quality to be used for D-MAD, or to automatically optimize the D-MAD threshold, the latter being investigated in section V.

According to Figure 6, the "Unified" OFIQ measure's quality scores appear to consistently fall with increasing degradation severity for all of the defect types.

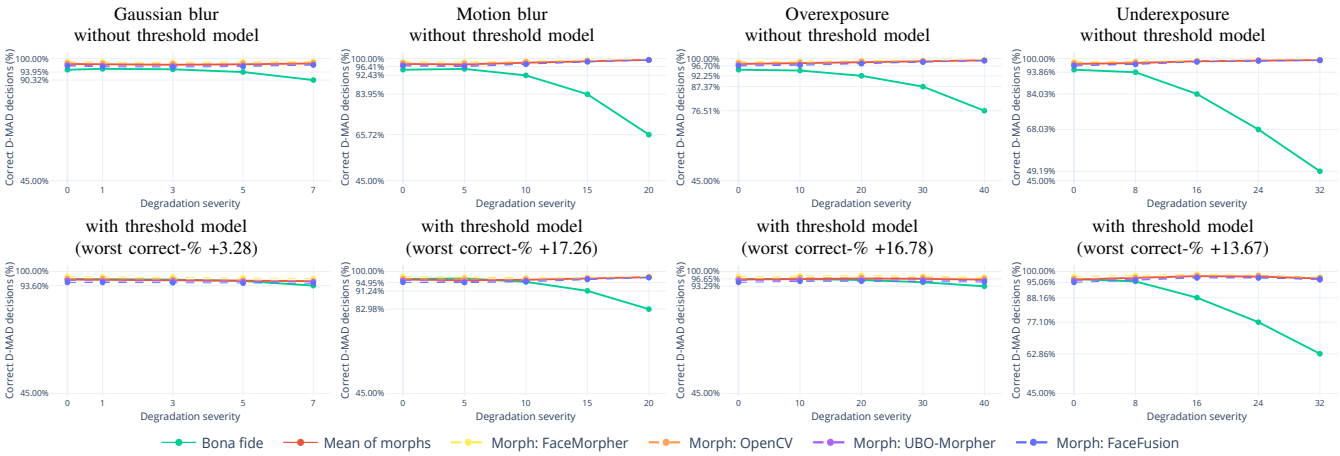

Fig. 7. A comparison between the percentage of correct D-MAD decisions without (top row) and with (bottom row) the D-MAD decision threshold optimization model from section V. The top row plots are analogous to the plots shown in Figure 4, except that only the test data subset from subsection V-A is used. The bottom row plots show results for the same data, but using the D-MAD decision thresholds from the model (instead of the constant 0.5 threshold). In the bottom row plot titles, the "worst correct-%" values all refer to the change of the bona fide Y-axis value for the highest degradation severity setting of the defect type (i.e. the worst correct-% in the top row plots).

As to be expected, the "Sharpness" measure is barely affected by the unrelated defect types "Overexposure" and "Underexposure", and the quality scores fall of consistently for the "Motion blur" defect type. For "Gaussian blur" the score distribution unexpectedly increases for the lowest non-zero degradation severity (1), before rather sharply falling to the lowest quality score value possible (0) for the higher severity settings. Notice that this response to the two blur types is converse to the previously shown D-MAD impact, where "Motion blur" clearly led to a stronger impact than "Gaussian blur".

The two OFIQ measures "Under-Exposure-Prevention" and "Over-Exposure-Prevention" are approximately unaffected by the two unrelated blur defect types, and the quality scores of both respond to their respective defect types "Underexposure" and "Overexposure", i.e. again as expected. A likely rather unimportant minor impact can also be observed on "Under-Exposure-Prevention" for "Overexposure" degradation, and vice versa on "Over-Exposure-Prevention" for "Underexposure" degradation, the former impact being a little more pronounced than the latter since the "Under-Exposure-Prevention" quality scores start at a slightly lower value for the non-degraded images.

V. D-MAD THRESHOLD OPTIMIZATION

A. Model training

Figure 1 illustrates the concept. For one given trusted live capture, the corresponding quality scores of the four examined OFIQ [16] measures are used as the input of the model. The output is the D-MAD score threshold, which can then be used instead of the D-MAD system's default threshold (0.5) to obtain a D-MAD decision from the D-MAD score.

The training setup's targeted output values are the optimal D-MAD score thresholds computed in the subsection IV-B analysis (Figure 5), for all the previously examined degradation configurations (including no degradation).

To train the model the dataset is randomly split into approximately 10% training data, 10% validation data, and 80% test data. The training data refers to the data with which a model is directly trained, and the performance of different model variants on the validation data is used to select one final model. Said validation data performance is computed in terms of the Root Mean Square Error (RMSE) between a model's predicted D-MAD threshold and a known targeted threshold optimum for all images in the validation data. This final model is then evaluated on the test data, analogously to the D-MAD decision evaluation in subsection IV-B (Figure 4).

Among the considered model variants, the one that yielded the best validation results was a "Histogram-based Gradient Boosting Regression Tree" model. More specifically, we used a Python implementation from the "scikit-learn" [19] package at version 1.2.2, namely "sklearn.ensemble.HistGradientBoostingRegressor", with the maximum number of iterations/trees set to 200.

Other considered model variants from the same Python package were "SVR", "DecisionTreeRegressor" (with AdaBoost), "RandomForestRegressor", "ExtraTreesRegressor", "KNeighborsRegressor", and "MLPRegressor", with multiple configuration parameter details that are omitted for brevity here. Various small custom artificial neural networks were also considered, using PyTorch [20] ("torch" package version 2.0.1). All of these other model variants did however yield worse validation data performance.

While the computational performance shouldn't be a concern for any of these model variants, the selected "HistGradientBoostingRegressor" model coincidentally also exhibits an especially good computational performance among them, with the training requiring less than one second (not including data loading/saving time), and the model predictions for the validation data performance computation requiring less than 100ms. In an operational scenario the computational requirements would thus predominantly stem from the D-MAD system and the OFIQ measures.

B. Model results

Figure 7 shows the results on the test data part of the dataset in terms of the correct D-MAD decision percentages with and without the model's threshold optimization. One the one hand, the results without the model use the default threshold, i.e. 0.5 as previously mentioned in subsection IV-B, analogously to the full dataset results in Figure 4. On the other hand, results that use the model for threshold optimization involve a different D-MAD score threshold for each trusted live capture, as explained in the prior subsection V-A.

Using the model's threshold optimization clearly improves the overall D-MAD decision performance across all defect types in this evaluation, despite the model only using four OFIQ measure quality scores as input. Since the percentages of correct bona fide D-MAD decisions was most affected by the degradation in this setup, they now benefit from the model's optimized thresholds. Conversely, the percentage of correct morph D-MAD decisions can be slightly decreased when the model is used. This is likely due to the training's focus on worst-case optimization (i.e. because the worst-case optimization thresholds from Figure 5 were used for training, as previously noted). Targeting different thresholds during training could alternatively avoid the slight impact on the correct morph D-MAD decision percentage at the cost of a slightly reduced bona fide decision improvement, if that were preferable for an operational setup. The same could also be achieved without training changes via the addition of a small offset constant to the model's threshold output.

VI. Conclusion

The experiments demonstrated that blur and exposure degradation can substantially affect D-MAD performance (subsection IV-B). Underexposure had a stronger D-MAD impact than overexposure (subsection IV-B), despite lower settings for the same degradation synthesis command type (Figure 2), and synthetic "motion blur" had a stronger impact than Gaussian blur (subsection IV-B) at roughly comparable settings (Figure 2). As expected, the unified OFIQ measure (MagFace [6]) responded to all defect types, whereas the response of the other examined OFIQ measures depended more on the defect type (subsection IV-C). Finally, a lightweight but effective D-MAD threshold optimization model was trained, using only used the examined OFIQ measures as input (section V).

Potential future works could for example examine other types of D-MAD and FIQA systems, other morphing algorithms, other forms of degradation, and other face image datasets in general. Future works could also focus in particular on the D-MAD threshold optimization, or they could examine the direct improvement of D-MAD models without the creation of separate models.

VII. Acknowledgements

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. Portions of the research in this paper use the FERET database of facial images collected under the FERET program, sponsored by the DOD Counterdrug Technology Development Program Office.

References

[1] ISO/IEC JTC1 SC37 Biometrics, *ISO/IEC 2382-37:2022 Information technology - Vocabulary - Part 37: Biometrics*, International Organization for Standardization, 2022.

[2] T. Schlett, C. Rathgeb, O. Henniger, J. Galbally, J. Fierrez, and C. Busch, "Face image quality assessment: A literature survey," *ACM Computing Surveys (CSUR)*, December 2021.

[3] A. Franco, M. Ferrara, C. Liu, C. Busch, and D. Maltoni, "On the impact of face image quality on morphing attack detection," in *Intl. Joint Conf. on Biometrics (IJCB)*. IEEE, 2024, pp. 1–9.

[4] B. Fu and N. Damer, "Face morphing attacks and face image quality: The effect of morphing and the unsupervised attack detection by quality," *IET Biometrics*, vol. 11, no. 5, pp. 359–382, 2022, https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/bme2.12094.

[5] B. Fu, N. Spiller, C. Chen, and N. Damer, "The effect of face morphing on face image quality," in *Intl. Conf. of the Biometrics Special Interest Group (BIOSIG)*, 2021, pp. 1–5.

[6] Q. Meng, S. Zhao, Z. Huang, and F. Zhou, "MagFace: A universal representation for face recognition and quality assessment," in *Conf. on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2021, pp. 14 225–14 234.

[7] G. Borghi, A. Franco, N. D. Domenico, M. Ferrara, and D. Maltoni, "V-MAD: Video-based morphing attack detection in operational scenarios," 2024, https://arxiv.org/abs/2404.06963.

[8] ImageMagick Studio LLC, "ImageMagick," 2024, https://imagemagick.org (version 6.9.11-60).

[9] U. Scherhag, C. Rathgeb, J. Merkle, and C. Busch, "Deep face representations for differential morphing attack detection," *IEEE Trans. on Information Forensics and Security*, 2020.

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

[11] P. Phillips, H. Moon, S. Rizvi, and P. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1090–1104, October 2000.

[12] P. Phillips, P. Flynn, T. Scruggs, K. Bowyer, J. Chang *et al.*, "Overview of the Face Recognition Grand Challenge," in *Conf. on Computer Vision and Pattern Recognition (CVPR)*, vol. 1. IEEE, June 2005, pp. 947–954.

[13] K. Raja, M. Ferrara, A. Franco, L. Spreeuwers, I. Batskos *et al.*, "Morphing attack detection - database, evaluation platform and benchmarking," *IEEE Trans. on Information Forensics and Security*, November 2020.

[14] J. Yang, P. Grother, M. Ngan, K. Hanaoka, and A. Hom, "Face analysis technology evaluation (FATE) part 11: Face image quality vector assessment," National Institute of Standards and Technology, NIST IR 8485 DRAFT SUPPLEMENT, April 2024.

[15] M. Ngan, P. Grother, K. Hanaoka, and J. Kuo, "Face analysis technology evaluation (FATE) part 4: MORPH - performance of automated face morph detection," National Institute of Standards and Technology, NISTIR 8292 DRAFT SUPPLEMENT, June 2024.

[16] Various contributors, "Open Source Face Image Quality (OFIQ) project," 2024, https://github.com/BSI-OFIQ/OFIQ-Project.

[17] J. Merkle, C. Rathgeb, B. Herdeanu, B. Tams, D. Lou, A. Dörsch *et al.*, "Open Source Face Image Quality (OFIQ) - Implementation and Evaluation of Algorithms," https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/OFIQ/Projektabschlussbericht_OFIQ_1_0.pdf, September 2024.

[18] Various contributors, "OFIQ Python adapter fork," 2024, https://github.com/torss/OFIQ-Project.

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[20] J. Ansel, E. Yang, H. He, N. Gimelshein *et al.*, "PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation," in *29th ACM Intl. Conf. on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24)*. ACM, Apr. 2024, https://pytorch.org/assets/pytorch2-2.pdf.