# DISTRIBUTED AND GDPR/IPR COMPLIANT BENCHMARKING OF FACIAL MORPHING ATTACK DETECTION SERVICES

Author list anonymized

*Author affiliations anonymized*

***Abstract:*** Having started in ANONYMIZED and lasting till ANONYMIZED, the research project ANONYMIZED, funded by ANONYMIZED, was the first inter-institutional research initiative established for designing Morphing Attack Detection (MAD) methods. The project is the prompt response to the paper "The Magic Passport" published by Ferrara et al. in 2014, demonstrating the threat that the face morphing attack (FMA) poses to the identity verification procedure based on facial photographs, including the usage of electronic Machine Readable Travel Documents (MRTD). Apart from the considerable scientific contribution reflected in a large number of international publications, an important result of the project is the distributed framework of MAD services, face image databases as well as the benchmarking service allowing for performing General Data Protection Regulation (GDPR) and Intellectual Property Rights (IPR) compliant evaluation. This paper introduces the MAD benchmarking framework by reporting its infrastructure, communication protocol and the results of an exemplary evaluation run. The design of the framework brought together the expertise of industrial companies with the innovative power of research institutions. The framework enables a statistically significant performance evaluation of MAD services. The individual MAD services as well as their combination may help countering the threat posed by FMA. The framework offers a research tool which could be used not only by project members but also by external parties.

***Keywords:*** Face Morphing Attack, Morphing Attack Detection, Distributed Benchmarking

## 1. INTRODUCTION

Biometric verification plays an increasingly important role in our daily life. Automated identity (ID) document checking technology increasingly supports border control officers and partly allows for automation of processes. Back in 2004, the International Civil Aviation Organization (ICAO) selected face as the primary biometric trait used with an MRTD. Along with all advantages of face verification there are serious security concerns caused by the vulnerability in the submission process of facial photographs [2]. If a morphed photograph arises in a document, both border guards and automated face recognition (AFR) systems are very likely to accept any of constituent individuals [3][8], abolishing the unique links between individuals and their ID documents. Although protection from FMA is a young research field, several research groups have already designed and prototypically implemented a bunch of morphing attack detection (MAD) approaches [9]. The still missing part of the research is the fair and GDPR/IPR compliant benchmarking of MAD approaches.

The trustworthiness of a benchmarking process highly depends on the experimental data. A benchmark maintainer should prepare two image sets: genuine face images and morphed face images. By collecting genuine face images, one should bear in mind that facial photographs, in particular those of a high quality, are regarded as personal biometric data which is protected by GDPR. Sharing of such data with third parties is prohibited by the European law warranting the image donors' right to request image removal at any moment. As a consequence, facial photographs must be stored on a protected media disabling the option of copying the data to any uncontrolled media. Putting the data into the public domain is not possible. An elegant solution is not to grant access to the database, but to ask the developers of MAD approaches to submit their algorithms for evaluation. It is important that the benchmarking is conducted by an independent body ensuring that the algorithms are not misused, e.g., disassembled or offered to third parties without consent of the owner. While preparing morphed face images, one should bear in mind that the morphing approach used by a perpetrator may differ from that used for generating experimental data.

While the number of publications on MAD approaches is growing fast (see recent overviews [9][16]), the efforts on the benchmarking of such methods are slow on the uptake. All former efforts on benchmarking of MAD approaches can be assigned to one of two categories: public benchmarks with the data unknown to MAD developers and individual benchmarks with self-collected or public data aiming at understanding the characteristics of MAD approaches and improving their performance.

Currently there exist two public benchmarks for MAD approaches: the FVC-onGoing Face Morphing Challenge maintained by the UNIBO [4] and the FRVT MORPH maintained by the NIST [11]. Both challenges provide i/o interfaces and encourage the potential participants to submit MAD solutions that are compliant to the given runtime environment specifications. The contributions are supposed to be executed on local servers of the organizers with undisclosed genuine and morphed face images and the benchmarking results are supposed to be publicly reported. At the time this paper was drafted (June 2019), the Web sites of both challenges reported no

participation results. The aforementioned public benchmarks share the same drawbacks. MAD algorithms have to be re-implemented to comply with the very restrictive run-time environment and additionally the numbers of submissions and test runs per participant are limited. Moreover, the composition of the test dataset cannot be influenced, which restricts the understanding of how specific image characteristics influence the error rates of a MAD algorithm. On the one hand, all these constraints make sense for a public benchmark, because otherwise the system would be prone to sensitivity attacks by participants aiming at dominating the challenge by creating detectors that perfectly fit to the test dataset instead of generally preventing FMA. On the other hand, to better understand the shortages of their MAD solutions, the researchers are forced to come up with alternative (non-public) benchmarking. Early scientific publications on specific aspects of benchmarking [5][6][12], were not implemented into a fully operational public benchmark.

Our proposed benchmarking framework is a network of RESTful Web services including those automatically generating high quality morphed images, MAD services, private face image databases, and a meta-database of image IDs. The benchmarking Web service, which is a core part of the framework, requests the image IDs that meet certain criteria from the meta-database. Then the images are derived from one of the image databases and sent to the MAD services. The responses of these MAD services are stored in the benchmarking log. A user communicates with the benchmarking Web service only by defining the criteria for image selection and by choosing the MAD services to test. All components of the framework and the communication between them remain hidden.

Experiments demonstrated the effectiveness of the benchmarking engine: We compared the performances of selected existing MAD services in an exemplary run on a particular dataset. The main benefits of such benchmarking are the secure transfer of private biometric images, the possibility to develop MAD solutions on different platforms using any programming language, the non-disclosure of the implementation details of MAD solutions, and the possibility to run an individually configured benchmark for any registered user. This design (together with a service-level agreement signed by the maintainers on the stateless nature of their services) ensures GDPR and IPR compliance. Since the MAD services are completely independent and maintained by different institutions, new services can be easily added to the framework. The same applies to image databases requiring only the registration of new images in the meta-database. The presented framework was designed and implemented within the research project ANONYMIZED enabling a fair comparison of MAD services provided by the project partners and a better understanding of advantages and shortcomings of certain MAD approaches.

## 2. METHODS

While designing the infrastructure for the benchmarking framework the focus was on security, sustainability and extendibility. The first core part of the framework is the infrastructure based on Representational State Transfer (REST) technology. The framework is in its nature a network of RESTful Web services which exchange data in a JSON format. The JSON objects are sent via HTTP POST requests. Web services are independent of each other so that they can be easily replaced or extended by the other ones. There are no restrictions on the hardware, operational system, or programming language for MAD algorithm providers. The code of Web services is undisclosed and the inter-service as well as client-service communication is done via the data exchange protocols.

The second core part is the secure storing and transfer of images. Datasets of face images are stored persistently on the servers of the image providers. The images are transferred via encrypted channels and the image consumers (e.g., MAD Web services) keep these images exclusively in a protected volatile memory. The images for benchmarking can be selected using a database containing image meta-data such as image ID, image type (genuine or morphed), image characteristics, and characteristics of a data subject. The annotations to morphed images include references to all images used for their generation.

The benchmarking framework architecture and the benchmarking workflow are illustrated in Figure 1. The *Benchmark Web Service* serves as the entry point for benchmarking requests. *Permitted Users* can initialize a benchmark by specifying a list of images or image meta-data, a list of MAD algorithms, and a list of image manipulation options. From the given image metadata, a request is constructed and sent to the *Image Meta-Database Web Service*. This meta database stores all meta data on the images hosted by the project members and links these images to unique identifiers. The result of a database request is merged with the explicitly provided image list. Each element of an image list contains the image ID as well as the ID of the image owner. Using this information several requests are constructed and sent to *Image Database Web Services* hosted by the corresponding vendors. The *Image Database Web Services* return the requested images to the *Benchmark Web Service*, which passes them to the image manipulation pipeline configured by the image manipulation options mentioned above. Afterwards, the images are sent to the selected MAD Web services (called *Algorithm_X Web Services*). The results of image evaluations are returned to the *Benchmark Web Service* and stored persistently.

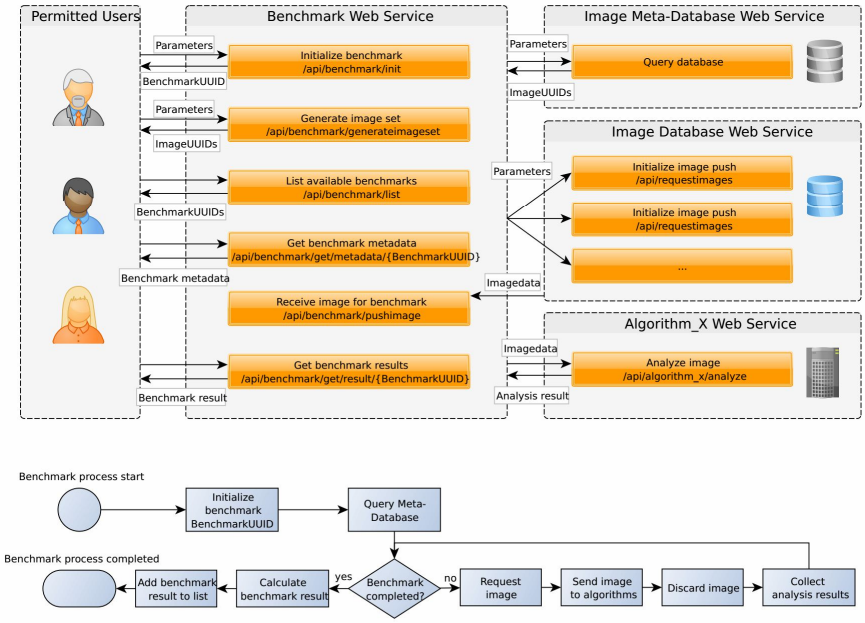



Figure 1: Architecture of the benchmarking engine and benchmarking workflow

The results of a benchmarking run can be requested via a unique BenchmarkUUID. These results can be reproduced by running a new benchmark with the same parameterization. In the case of several benchmarks running in parallel and operating with same images, the analysis requests are sent only once to the corresponding MAD web services and the evaluation results are used for all benchmarks. The MAD services support both "blind" detection based on a passport photograph only and detection in the presence of a "live" photograph. Thanks to the integrated image manipulation engine, the influence of anti-forensic approaches to the detection performance can be evaluated. The interface based on Web services enables easy integration with the Automated Border Control (ABC) reference environment of Bundesdruckerei GmbH, Germany. The list of Web service providers is given in Table 1. Note that Table 1 also includes the morphing Web services which are strictly speaking not a part of the benchmarking framework. However, these Web services are used to fill the image databases with morphed face images.

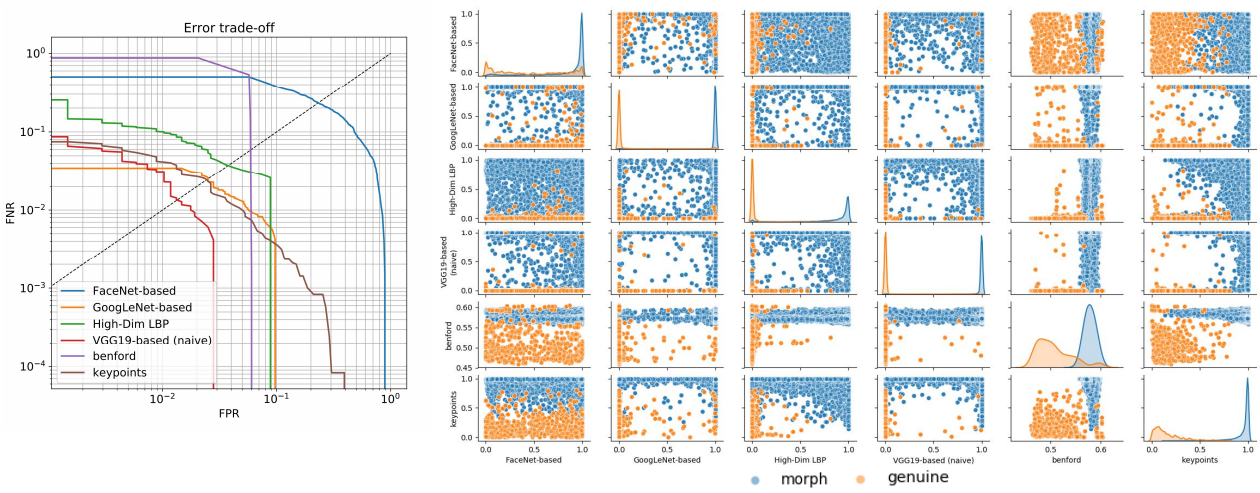
Table 1: Hosts of Web services

| | |
|---|---|
| Benchmark Web service | - ANONYMIZED |
| Meta-Database Web service | - ANONYMIZED |
| Image Database Web services | - ANONYMIZED<br>- ANONYMIZED<br>- ANONYMIZED<br>- ANONYMIZED |
| Algorithm_X Web services | - ANONYMIZED<br>- ANONYMIZED<br>- ANONYMIZED |
| Morphing Web services | - ANONYMIZED<br>- ANONYMIZED<br>- ANONYMIZED |

By May 2019, the total number of face images registered in the meta-database exceeded five million. Currently, ten MAD algorithms are available: four hosted by ANONYMIZED, four by ANONYMIZED, and two by ANONYMIZED.

## 3. FINDINGS AND ARGUMENT

Here, we report the results of an exemplary benchmarking run lasting from November 1st to December 3rd, 2018. Based on 680 genuine images (605 male/75 female) selected from the PUT face database [13] we generated 12000 morphed face images with two algorithms, 6000 each. Both morphing algorithms are deployed as Web services, the first one by ANONYMIZED and the second one by ANONYMIZED. There are 5321/679 male/female morphs generated by ANONYMIZED Web service and 5730/270 male/female morphs by ANONYMIZED Web service. The ethnicities of the data subjects are Caucasian, Latin, and Middle Eastern. For the purpose of quality assessment, all morphed face images were compared with constituent face images using the Dermalog Face Recognition software [1] as a commercial off-the-shelf product. The requirement for the inclusion of a morphed face image into the dataset is that both comparison scores exceed 80% similarity.




(a) (b)

Figure 2: (a) DET curves and (b) discriminatory power of the MAD Web services

We benchmarked six MAD algorithms: *FaceNet-based* [17], *GoogLeNet-based* [14], *High-Dim LBP* [17], *VGG19-based (naive)* [15], *benford* [10], and *keypoints* [7]. For performance evaluation, we used standard metrics for binary classification problems: FPR vs. FNR. Note that since we endeavour to detect morphs, morphed images are considered as positive samples and genuine images as negative. FPR is the fraction of genuine images that are falsely classified as morphed images and FNR is the fraction of morphed images that are falsely classified as genuine images. The Detection Error Trade-off (DET) graph in Figure 2(a) demonstrates the performances of the MAD algorithms at different operating points. The discriminatory power of the evaluated algorithms is visualized in Figure 2(b). The diagrams on the main diagonal show the matching score distributions of morphed (blue) and genuine (orange) samples while non-diagonal diagrams reveal the potential for fusion by demonstrating the pair-wise correlation between MAD algorithms. A point on a non-diagonal diagram represents an image by a pair of matching scores resulting from two different MAD algorithms (one on the X-axis and another on the Y-axis). A matching-score fusion of MAD algorithms is expected to improve the recognition performance if the blue and orange points can be separated by a diagonal line. Horizontal or vertical separation lines indicate the domination of one MAD algorithm over another and a limited potential for fusion. Note that the MAD services operate at fixed thresholds, i.e. whether the algorithm classifies an image as morphed or genuine critically depends on the chosen decision boundary.

Table 2: Detection performance of the MAD Web services, the best performances are highlighted

| | FPR | FNR | FNR @ 0.01% FPR | FNR @ 0.1% FPR | FNR @ 1% FPR | FNR @ 10% FPR | EER |
|---|---|---|---|---|---|---|---|
| *FaceNet-based* | 43.53% | 12.54% | 59.23% | 59.09% | 57.69% | 37.65% | 23.21% |
| *GoogLeNet-based* | 3.97% | 1.52% | **4.06%** | **4.03%** | 3.66% | 0.30% | 2.53% |
| *High-Dim LBP* | 0.15% | 18.02% | 25.96% | 25.53% | 10.03% | 1.45% | 3.77% |
| *VGG19-based (naive)* | 1.03% | 2.75% | 10.29% | 9.24% | **3.13%** | **0.00%** | **1.36%** |
| *benford* | 53.09% | 0.00% | 99.94% | 99.38% | 93.82% | 0.00% | 5.97% |
| *keypoints* | 4.71% | 1.14% | 7.93% | 7.66% | 4.19% | 0.36% | 2.43% |

Table 2 shows the factual FPR and FNR values with the predefined thresholds used by MAD services, theoretical FNR values at particular levels of FPR, and the Equal Error Rates (EER). Regarding the EER and the detection performance at FPR higher than 1%, the best algorithm is *VGG19-based (naive)* followed by *GoogLeNet-based* and *keypoints*. At low levels of FPR (0.1% and lower), the *VGG19-based (naive)* has FNR of over 9%, while the *GoogLeNet-based* of around 4% and *keypoints* of under 8%. The error rates of *High-Dim LBP*, *benford* and *FaceNet-based* are strongly imbalanced requiring more careful selection of decision thresholds. Observing the EER values, we see that with properly selected decision thresholds the MAD services demonstrate solid detection performance. However, these error rates are too high clearly indicating that the algorithms are still not mature for practical application.

## 4. CONCLUSIONS

Thanks to its design based on the REST technology, the presented benchmarking framework is a powerful and flexible tool for GDPR/IPR compliant evaluation of MAD approaches. Our proposed interfaces for MAD services support both "blind" detection based on a passport photograph only and detection in the presence of a "live"

photograph. The integrated image manipulation tools enable for evaluation of the influence of anti-forensics. The benchmarking results are reproducible and transparent for the benchmark users.

The demonstrated benchmarking run does not cover all capabilities of the framework, but gives an idea how the benchmark can be configured and how the results can be visualized.

Due to the flexibility of interfaces, the MAD services can be used not only as a part of the framework, but also individually. Currently, the MAD Web services are integrated into the ABC reference environment of Bundesdruckerei GmbH, Germany.

The proposed benchmarking framework can be used as an alternative to FVC-onGoing Face Morphing Challenge and NIST FRVT MORPH having an advantage of providing more flexibility by developing of MAD approaches and granting more transparency in test image datasets. Parties interested in benchmarking MAD approaches are invited to contact the authors in order to register as users of the framework so that they can browse through existing benchmarks as well as configure and run own ones.

## 5. REFERENCES


1. Dermalog Face Recognition, https://www.dermalog.com/products/software/face-recognition, online, 06.06.2019
2. M. Ferrara, A. Franco, and D. Maltoni, “The Magic Passport,” Proc. IEEE Int. Joint Conf. on Biometrics, pp. 1–7, 2014
3. M. Ferrara, A. Franco, and D. Maltoni, “On the Effects of Image Alterations on Face Recognition Accuracy,” in Face Recognition Across the Electromagnetic Spectrum, T. Bourlai (Ed.), pp. 195–222, Springer, 2016
4. FVC onGoing: Face Morphing Challenge, https://biolab.csr.unibo.it/FVCOnGoing/UI/Form/BenchmarkAreas/BenchmarkAreaFMC.aspx, online, 06.06.2019
5. M. Gomez-Barrero, C. Rathgeb, U. Scherhag, and C. Busch, “Predicting the vulnerability of biometric systems to attacks based on morphed biometric information,” IET Biometrics 7(4): 333-341, 2018
6. M. Hildebrandt, T. Neubert, A. Makrushin, and J. Dittmann, “Benchmarking Face Morphing Forgery Detection: Application of StirTrace for Impact Simulation of Different Processing Steps,” Proc. 5th Int. Conf. on Biometrics and Forensics (IWBF), 2017
7. ANONYMIZED, 2017.
8. A. Makrushin, T. Neubert and J. Dittmann. “Automatic generation and detection of visually faultless facial morphs,” Proc. 12th Int. Joint Conf. on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 6: VISAPP, pp. 39–50, 2017
9. A. Makrushin and A. Wolf, “An Overview of Recent Advances in Assessing and Mitigating the Face Morphing Attack,” Proc. 26th European Signal Processing Conference (EUSIPCO), pp. 1017-1021, 2018
10. ANONYMIZED, 2018
11. NIST FRVT MORPH, https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt-morph, online, 06.06.2019
12. T. Neubert, A. Makrushin, M. Hildebrandt, C. Kraetzer and J. Dittmann, “Extended StirTrace Benchmarking of Biometric and Forensic Qualities of Morphed Face Images,” IET Biometrics 7(4):325–332, 2018
13. A. Kasiński, A. Florek, and A. Schmidt, “The PUT face database,” Image Processing and Communications 13:59-64, 2008
14. ANONYMIZED, 2017
15. ANONYMIZED, 2018
16. U. Scherhag, C. Rathgeb, J. Merkle, R. Breithaupt, and C. Busch, “Face Recognition Systems Under Morphing Attacks: A Survey,” IEEE Access 7:23012-23026, 2019
17. ANONYMIZED, 2018