h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES

Hochschule Darmstadt

Fachbereich Informatik
Fachbereich Mathematik &
Naturwissenschaften

Analysis of face embeddings to facilitate image pre-selection for face morphing

Abschlussarbeit zur Erlangung des akademischen Grades
Master of Science (M.Sc.)

vorgelegt von

Roman Kessler
Matrikelnummer: 764734

Referent	:	Prof. Dr. Christoph Busch
Korreferent	:	Dr. Juan Tapia Farias
Externer Betreuer	:	Dr. Kiran Raja
Ausgabedatum	:	11.10.2021
Abgabedatum	:	17.01.2022

ERKLÄRUNG

Ich versichere hiermit, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die im Literaturverzeichnis angegebenen Quellen benutzt habe.

Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten oder noch nicht veröffentlichten Quellen entnommen sind, sind als solche kenntlich gemacht.

Die Zeichnungen oder Abbildungen in dieser Arbeit sind von mir selbst erstellt worden oder mit einem entsprechenden Quellennachweis versehen.

Diese Arbeit ist in gleicher oder ähnlicher Form noch bei keiner anderen Prüfungsbehörde eingereicht worden.

Darmstadt, 17.01.2022

Roman Kessler

ABSTRACT

Face Morphing Attacks pose a novel threat to the security of identification documents. The fusion of the face images of two or more – similarly looking – individuals during the application process for a new travel document (i.e., passport) or identity card enables both individuals to travel with the same document. In order to develop algorithms to detect morphing attacks, large data sets of morphed face images are needed, for which in turn many similarly looking individuals need to be paired.

The study at hand uses face embeddings of openly accessible face recognition models to describe similarity between individuals. It aims at finding appropriate face recognition models, metrics to quantify similarity, morphing algorithms to fuse facial images of paired individuals, and soft biometric characteristics to analyze the attack potential of face morphs.

Results demonstrate, that image pre-selection based on Cosine or Euclidean distances between face embeddings highly improves the attack potential of morphs. Especially ArcFace and MagFace provide valuable face embeddings to quantify similarity for pre-selection. Both open source, as well as Commercial Off-The-Shelf Face Recognition Systems get fooled by morphed faces. Landmark-based, closed source morphing algorithms pose high risk for any of the tested Face Recognition Systems. On the other hand, MagFace embeddings further emerge as valuable means to detect morphed face images. Soft biometrics characteristics however were only partially relevant to predict morph success, if morphing has been conducted within similar age, gender, and race groups.

The results emphasize that face embeddings are valuable instruments on both sides of the morphing attack, image pre-selection for face morphing and detection of morphed faces.

ZUSAMMENFASSUNG

Gesichtermorphing Angriffe stellen eine neue Gefahr für die Sicherheit von Idenditätsnachweisen dar. Die Verschmelzung zweier – sich ähnlich sehender – Lichtbilder zu einem Morph, der in der Antragsstellung für ein Identitätsdokument (Pass, Personalausweis) eingereicht wird, ermöglicht es beiden Beteiligten gleichermaßen mit dem ausgestellten Dokument zu reisen. Um Algorithmen für die Erkennung solcher Morphing Angriffe zu entwickeln werden große Mengen von Morphs benötogt, welche wiederum aus vielen – sich ähnlichen – Gesichter-Paaren zusammengesetzt sein müssen.

Die hier vorliegende Studie benutzt Gesichter-Embeddings von Open Source Gesichtserkennungs-Modellen um Ähnlichkeit zwischen Individuen zu beschreiben. Das Ziel ist, passende Gesichtserkennungs-Modelle, Ähnlichkeitsmaße, Morphing-Algorithmen, und Soft-Biometrische Eigenschaften zu analysieren, um das Angriffspotential von Morphs zu verbessern.

Die Ergebnisse zeigen, dass wenn die Cosinus-Distanz oder die Euklidische Distanz zwischen zwei Gesichtern als Ähnlichkeitsmaß für die Paarung von Lichtbildern genommen wird, das Angriffspotential der resultierenden Morphs erhöht wird. Speziell ArcFace und MagFace stellen geeignete Gesichter-Embeddings für die Berechnung dieser Ähnlichkeit bereit. Sowohl Open Source, als auch kommerzielle Gesichtserkennungssysteme werden von den resultierenden Gesichtermorphing Angriffen getäuscht. Landmarkbasierte, nicht-öffentliche Morphingalgorithmen generieren hochwertige Morphs, welche ein hohes Risiko für die getesteten Gesichtserkennungssysteme darstellen. Andererseits stellen sich vor allem MagFace Gesichter-Embeddings als nützliches Werkzeug für die Erkennung von gemorphten Gesichtern heraus. Soft-Biometrische Eigenschaften sind nur zum Teil für den Erfolg des Morph-Angriffs ausschlaggebend, zumindest wenn innerhalb bestimmter Alter-, Geschlechts- und Ethnizitätsgruppen gemorpht wurde.

Die Ergebnisse betonen die Wichtigkeit von Gesichter-Embeddings auf beiden Seiten des Angriffs. Zum einen können sie für die Paarfindung vor dem Morphing eingesetzt werden, zum anderen wiederum für die Erkennung von gemorphten Gesichtern.

CONTENTS

LIST OF ABBREVIATIONS

APCER Attack Presentation Classification Error Rate

BPCER Bona fide Presentation Classification Error Rate

$BPCER_{10}$ BPCER at which the APCER is 10% (= 0.1)

COTS Commercial Off-The-Shelf

DET Detection Error Trade-off

ECDF Empirical Cumulative Distribution Function

FMR False Match Rate

FNMR False Non-Match Rate

$FRGCv_2$ Face Recognition Grand Challenge database version 2

FRS Face Recognition System

GAN Generative Adversarial Network

ICAO International Civil Aviation Organization

MAD Morphing Attack Detection

S-MAD Single-image Morphing Attack Detection

D-MAD Differential-image Morphing Attack Detection

MIPGAN Morphing through Identity Prior driven Generative Adversarial Network

mRMR minimal Redundancy, Maximal Relevance

MMPMR Mated Morph Presentation Match Rate

prodAvgMMPMR product Average Mated Morph Presentation Match Rate

MVR Morph Vulnerability Rate

RMMR Relative Morph Match Rate

RMSE Root Mean Squared Error

SVM Support Vector Machine

TMR True Match Rate

Part I

THESIS

1 INTRODUCTION

1.1 MOTIVATION

Automated biometric recognition plays an integral role in settings of access control, criminal investigation, and surveillance [Kor+20]. In particular for automated border control, the observation and analysis of the face characteristics is becoming increasingly important for identity verification [Rio+16]. For instance, to support immigration officers at borders or airports, automated Face Recognition Systems (FRSs) increase the throughput of travelers and similarly reduce costs.

In a typical identity verification process, a biometric reference image – i.e., a passport photograph – is compared to one or multiple probe images – i.e., live photographs taken at the border. A similarity score is then computed between reference and probe image, and the subject might cross the border if their similarity score exceeds a certain threshold.

The operation of an automated FRS requires a particular security of the system. Security however can be compromised by means of a so-called morphing attack [Sch+16]. In a morphing attack, the face images of two or more subjects are combined to form a morphed face image (see e.g., Fig. 7). This morphed face image serves as biometric reference for the FRS, and is stored for instance in the passport. The calculated similarity score between the morphed face reference image and a bona fide probe image can be high enough to exceed some decision threshold τ, resulting in a successful verification of identity. As a consequence, two or more individuals may use the same passport for border crossing, and a one-to-one association between a passport and an individual is broken.

A significant research area is the so-called Morphing Attack Detection (MAD), in which algorithms are created or trained to recognize a morphing attack [Ven+21]. These algorithms are often based on machine learning. Therefore, an enormous amount of data for training is required (e.g., [FFM21]). However, high quality morphs often need manual post-processing to reduce artifacts [Sch+16; Sch+17], limiting the amount of high quality morphs available for large-scale training of MAD algorithms.

For these reasons, it is important to develop criteria which allow for an informed selection of two (or more) individuals, suitable to create a high quality morph image [RSB20], which does not – or to a less extent – rely on manual postprocessing. These criteria can then be used to find large numbers of possible pairs of suitable source images, from which morphs can be generated in an automated fashion, and a database of morphed images can be created for future research on MAD.

1.2 PREVIOUS WORKS

Previous research has shown, that an adequate pre-selection of possible morph pairs can diminish two things: First, the choice of the morphing algorithm applied is of less relevance [RSB20]. Second, the amount of artifacts produced by an automated morphing algorithm is reduced, rendering a FRS more vulnerable to the morph attack [RSB20]. A large database of morphed images not only allows for better training and testing of MAD algorithms. It further allows for an statistical analyses of the performances of FRSs, and may ultimately lead to a better understanding of the image properties which are predictive for the success of a morph attack.

For manual image pre-selection, some heuristic criteria have been applied in the past. For instance, soft biometrics characteristics have been used, to morph only subjects of similar age, same gender, or same race [Rag+17; Raj+20; Sch+17]. Further, other characteristics such as shape of hair, skin tone, differences in landmark position and euclidean distance between *face embeddings* extracted from the OpenFace model [ALS] have shown positive effects on the performance of a morph [RSB20].

1.3 FACE EMBEDDINGS AS A KEY FOR IMAGE PRE-SELECTION

In the study at hand, feature embeddings are used to perform image pre-selection. Feature embeddings are low-dimensional representations of high-dimensional input images, such as faces [Wil18]. In the context of face recognition termed *face embeddings*, feature embeddings are point representations in latent space learned during the training of a face recognition neural network [SJ19]. The distance between two face embeddings – for instance, the Euclidean distance between them – directly corresponds to the similarity of the faces [SKP15]. Therefore, for face identity verification, one can calculate distances on the embeddings and apply some threshold τ on the distance value (or accordingly on the similarity value) to decide whether the two face samples originate from the same identity or not.

The general idea is the following: a small distance between two subjects' face embeddings corresponds to a high similarity perception in a human observer. In turn similar looking faces could be able to produce more realistic morphs compared to a combination of two facial images which do not look particularly similar. Therefore, instead of manually handcrafting face pairs from a large database, choosing corresponding faces based on the similarity of their embeddings can automate the pair selection process. Further, automating the pair selection process (i.e., pre-selection) renders it comprehensible, reproducible, scalable, and lastly less subjective than manual approaches.

1.4 FACE RECOGNITION, MORPHING, AND ATTACK DETECTION

A variety of face recognition models based neural networks or convolutional neural networks have been published during the last years (e.g., [Tai+14; PVZ15; Den+19; Men+21]). These networks produce different embeddings for the same face image. Consequently, the distances between the face embeddings, and the pair selection of face images will be different, depending on which face recognition model was used to extract the embeddings.

Furthermore, different face morphing algorithms have been published (e.g., [Que; FFM14; Zha+21; Raj+20; Rag+17]). The majority of algorithms is landmark-based [KS14; Que; FFM14; Raj+20; Rag+17], fusing the raw faces images to an average face based on estimated facial landmarks. Others are based on Generative Adversarial Networks (GANs), morphing in latent space and generating the resulting morph from the latent space [Zha+21].

Both the choice of FRS for image pre-selection, and the FRS for verification of the morph can exert an influence on the success of the morph, i.e., its success to fool the verification FRS. Likewise, the morphing algorithm used can have a similar effect on the success of the morph, as well as the distance metric used for image pre-selection.

The embeddings of a particular FRS have already been tested alongside with other characteristics to enhance image pre-selection [RSB20]. The study used OpenFace [ALS] to extract face embeddings for image pre-selection and compared this pre-selection method to alternative pre-selection methods based on e.g., several soft biometrics or visual descriptors. Pre-selection based on OpenFace embeddings outperformed the alternatives in terms of fooling a FRS to generate higher similarity measures between morphs and respective bona fide images, and the morph to escape the detection of a MAD algorithm [RSB20].

1.5 AIMS OF THE PRESENT STUDY

The study at hand aims at an excessive analysis of image pre-selection based on embedding vectors.

- First, different distance metrics for image pre-selection – applied on the face embeddings – are evaluated. Therefore, the application of Euclidean distance, Cosine distance, and a similarity metric based on mutual information are tested. Pairs are selected based on low distance or high similarity, respectively. The resulting pairs are morphed, and the resulting morphs be evaluated based on the vulnerability they pose to a FRS, i.e., how likely a FRS is successfully tricked by the morphed image.
- Second, face embeddings can be produced by different FRSs, and new evolutions of FRSs are published frequently. Therefore, several state-of-the-art FRSs are deployed to generate the embeddings which are then used for image pre-selection. Further, the very same FRSs are also exam-

ined for identity verification of the morphs, alongside with two Commercial Off-The-Shelf (COTS) FRSs.

- Third, different morphing algorithms can be used to generate the morphs. The present study examines how different morphing algorithms contribute to the success of a face morph attack based on pre-selected images.
- Fourth, the resulting morphed images are evaluated using a MAD algorithm to assess if pre-selection is beneficial to exceed a better recognition score and reduce the probability that a MAD algorithm successfully detects the morphed image.
- Fifth, randomly morphed images are analyzed to disentangle factors increasing the success probability – in terms of fooling an FRS – of a morph attack.
- Finally, embeddings of soft biometrics models are tested on their ability to predict the success of a morph attack. Separate soft biometrics models are applied for gender, age, and race to evaluate if the similarities between subjects in those soft biometrics embeddings contributes to the success of a morph attack.

2 METHODS

To answer the current research questions, three general analysis tracks were constructed (Tab. 1). The tracks differed with respect to the underlying face data sets (Section 2.2), the FRSs used to extract face embeddings for image pre-selection (Section 2.3), the distance metrics applied for pairing (Section 2.5), the morphing algorithms used (Section 2.7), the FRSs used for verification (Section 2.3), or the application of MAD (Section 2.11).

This section is structured as follows: First, the coarse processing tracks will be summarized (Section 2.1). After that, more detailed descriptions of the single processing steps follow (Sections 2.2–2.11).

	track I	track II	track III
data set	FRGCv2	UNCW	UNCW
pre-selection embeddings	ArcFace	ArcFace DeepFace VGG-Face MagFace	age model (VGG) gender model (VGG) race model (VGG)
pre-selection distance\ similarity	Euclidean Cosine mutual information	Cosine	element-wise -Euclidean
morphing algorithm	Alyssaq UBO NTNU	Alyssaq UBO NTNU MIPGAN	Alyssaq
verification FRS	ArcFace	ArcFace DeepFace VGG-Face MagFace COTS1 COTS2	ArcFace
MAD embeddings		ArcFace MagFace	

Table 1: Different analysis tracks were used in the present study. The tracks are displayed in a vertical direction. Each track varied in the data sets used, the FRSs used to extract embeddings for image pre-selection, the distance metrics applied for pairing, the morphing algorithms used, the verification FRS used, or if MAD was applied.

2.1 PROCESSING TRACKS

2.1.1 *Track I*

Track I (Tab. 1, Fig. 1) was conducted to find an suitable distance metric (or similarity metric) for image pre-selection. A suitable metric for image pre-selection should therefore increase the mated morph verification rates of the resulting morphs. Therefore, the rather small FRGCv2 data set (Section 2.2.1) has been analyzed, using embeddings of the ArcFace model (Section 2.3.3) for image pre-selection. Three different distance (or similarity) measures have been calculated, Euclidean distance (Section 2.5.1), Cosine distance (Section 2.5.2), or Mutual Information as measure of similarity (Section 2.5.3). After pairing the images based on low distance (or high similarity), three different morphing algorithms have been deployed, Alyssaq morpher (Section 2.7.1), UBO morpher (Section 2.7.2), and NTNU morpher (Section 2.7.3). All morphs were verified against bona fide probe images of each subject using ArcFace (Section 2.3.3).

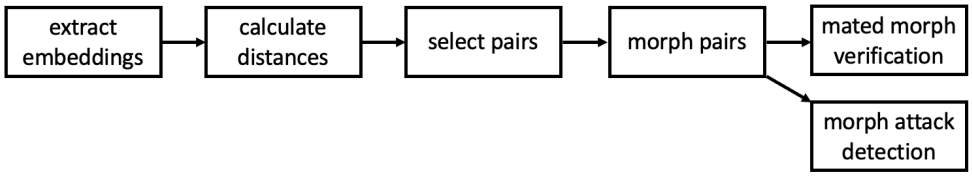

Figure 1: General workflow for track I and track II. The displayed processing steps were performed using different FRS to extract embeddings, different distances, different morphing algorithms, and different FRS for mated morph verification. See Table 1 for more information. In addition, track II included a MAD step.

2.1.2 *Track II*

Track II (Tab. 1, Fig. 1) was primarily conducted to evaluate the FRSs from which the embeddings were retrieved for image pre-selection. For this, the larger UNCW database was used (Section 2.2.2). Only Cosine metric was applied to determine the distances between the single face images (Section 2.5.2), since this metric performed best in Track I (Section 3.1). Morphing has been conducted using Alyssaq FaceMorpher (Section 2.7.1), UBO morpher (Section 2.7.2), NTNU morpher (Section 2.7.3), and additionally using MIPGAN (Section 2.7.4). Image verification has further been conducted with several FRSs, ArcFace (Section 2.3.3), DeepFace (Section 2.3.1), VGG-Face (Section 2.3.2), MagFace (Section 2.3.4), and two COTS algorithms termed COTS1 and COTS2 (Section 2.3.5). Furthermore, the morphs of Track II were evaluated using a Differential-image Morphing Attack Detection (D-MAD) algorithm (Section 2.11).

2.1.3 *Track III*

Track III (Tab. 1, Fig. 2) aimed at evaluating if image pre-selection can be done based on the embeddings of soft biometrics models. Therefore, the larger UNCW database has been used (Section 2.2.2). Embeddings were extracted using an age, gender, and race models (Section 2.4). All three models were derived from the VGG-Face model (Section 2.3.2). Euclidean distances have been calculated element-wise between the embeddings of the two bona fide images (Section 2.4). All pairs have been morphed using Alyssaq morpher (Section 2.7.1), because of its ability to conduct a large amount of morphs with low computational burden. Morphs have been verified using ArcFace (Section 2.3.3). In this track, a feature selection approach in combination with a regression model was deployed to evaluate if a subset of embeddings was able to predict the success of a morph attack (Section 2.12).

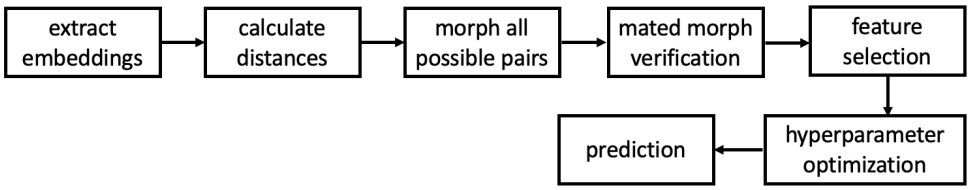

Figure 2: General workflow for track III. Contrary to track I and track II, the single processing steps were only performed once. However, all possible morph pairs were created, resulting in an extensive database of morphed images.

In the following, the respective data sets, FRSs, the calibration of FRSs, distance metrics, pre-selection, feature selection algorithms, face morphing algorithms will be introduced, alongside the MAD procedure, statistical methods and single processing steps.

2.2 FACE IMAGE DATA SETS

Within the course of this study, two different face image data sets were used for different purposes.

2.2.1 *FRGCv2*

The Face Recognition Grand Challenge database version 2 (FRGCv2) [Phi+05] database originated from a data science challenge and originally comprised several thousand face images. Here, an in-house built subset of face images was used. The face images in this data set were already the result of a handcrafted pre-selection procedure. Further, several processing steps were conducted for the images to align with the International Civil Aviation Organization (ICAO) standard for machine-readable travel documents [Int15]. Therefore the presently used FRGCv2 data set comprised around 70 data subjects with around 10 face captures each. Morphing was conducted between morphs of equal gender.

2.2.2 *UNCW*

The UNCW-MORPH face data set (academic version, short: UNCW) is distributed by the face aging group of University of North Carolina Wilmington (UNCW) [NCW]. It comprised over 55,000 face images of more than 13,000 data subjects, captured between 2003 and 2007. Contrary to its name, the UNCW database contained bona fide face images rather than morphed images. The images show frontal faces with (largely) neutral expressions, rendering the data set viable for face morphing. The image resolution was between 200×240 px and 400×480 px. Each image was accompanied by labels for exact age, gender (binary), and race (e.g., white, black, hispanic).

Figure 3: Samples from two exemplary data subjects from the UNCW data set. Each row illustrates five captures from a particular data subject. Captures were retrieved over a period of up to four years. Adapted from [UNC].

The raw data set was processed as follows: First, all samples were checked for neutral facial expressions. Therefore, the emotion model of the LightFace package [SO20] was used. All samples, for which "neutral" was not the dominant emotion, or samples, where the emotion model failed, were discarded from morphing. Afterwards, all subjects with less than five remaining samples were dismissed, as four bona fide probe samples are needed for vulnerability analysis illustrated in Figure 12. For the remaining data subjects, the first sample (in chronological order) was used for morphing. Out of the 55,134 samples from 13,618 data subjects of the raw data set, 22,992 samples from 3,337 data subjects remained in the data set.

Morphing then was conducted based on several criteria. Each candidate pair (based on their distances) was evaluated based on the labels of the soft biometrics age, gender, and race. A pair was only morphed if they corresponded in gender and race affiliations. Further, a pair was only morphed if the age difference of the two data subjects was less or equal to 5 years.

FRS	# of embeddings
ArcFace	512
DeepFace	4096
VGG-Face	2622
MagFace	512
age (VGG)	4096
gender (VGG)	4096
race (VGG)	4096

Table 2: The number of embeddings per FRS.

2.3 FACE RECOGNITION SYSTEMS

Different state-of-the-art implementations of FRSs have been used to extract either face embeddings for image pre-selection or for biometric verification. ArcFace (Section 2.3.3), VGG-Face (Section 2.3.2), DeepFace (Section 2.3.1), and MagFace (Section 2.3.4) were deployed for pre-selection based on face embeddings. The lengths of the embeddings vectors are illustrated in Table 2). The same FRSs have been used for verification, alongside with two COTS FRSs (Section 2.3.5).

For ArcFace, VGG-Face and DeepFace (Facebook), Tensorflow implementations of the respective models have been used which were embedded in the software distribution of the LightFace[1] repository [SO20]. Faces (bona fide and morphed faces) were aligned and cropped using *mtcnn* [Zha+16] and rescaled to 112×112 px before validation. Lastly, the respective FRSs returned different numbers of face embeddings, which are illustrated in Table 2.
In the following, face recognition models are listed based on their publication dates.

2.3.1 *DeepFace*

DeepFace was developed and published 2014 by researchers of Facebook [Tai+14]. Like all other open source FRSs used in the present study, DeepFace is based on convolutional neural networks [AMAZ17] trained with a large scale face image database. At the time of publication, it outperformed most state-of-the-art FRSs and nearly reached human-level performance in face recognition [Tai+14].

1 The LightFace repository [SO20] is also called DeepFace repository, but is not to be confused with the DeepFace model [Tai+14] which was developed by Facebook. The repository is not related to Facebook, but includes an implementation of its DeepFace model alongside other face recognition models.

2.3.2 *VGG-Face*

VGG-Face[2] was published in 2015 [PVZ15]. It outperformed the performance of Facebook's DeepFace model, and was further trained on an openly accessible face data set, contrary to the DeepFace model. The model's coarse architecture is depicted in Figure 4.

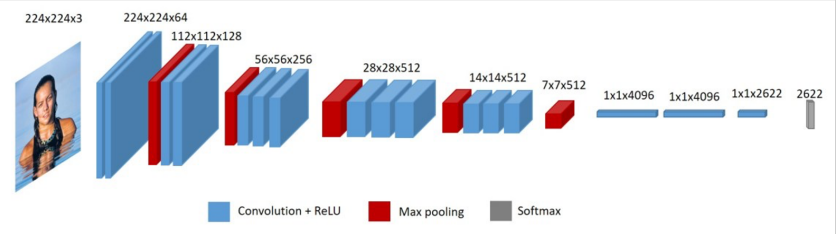

Figure 4: VGG-Face model. Image retrieved from https://sefiks.com/2018/08/06/deep-face-recognition-with-keras/. The input face image is propagated through a network consisting of several convolutional and max-pooling layers, followed by several fully connected layers. The output of a the network is a Softmax layer with 2622 neurons. The activations of those neurons correspond to the embeddings used to describe the face and to verify the identity by applying a distance function. The VGG-Face model was further used as baseline architecture for models estimating soft biometrics (see Section 2.4).

2.3.3 *ArcFace*

ArcFace was published 2019 and since than counts as state-of-the-art model for face recognition [Den+19]. The main contribution of the ArcFace model was the "Additive Angular Margin Loss", which has been shown to enhance discriminate power of the feature embeddings [Den+19]. ArcFace outperformed competing open source face recognition models at the time of publication.

2.3.4 *MagFace*

In their recent work, Meng et al. refined the ArcFace loss in a way to incorporate image quality in the model's learning process [Men+21]. MagFace learns the distribution of a class (i.e., identity) by locating the high quality samples in the center of the distribution and the low quality samples at its margins. MagFace slightly outperformed the ArcFace algorithm [Men+21] at the time of publication.

2 The titel of the paper is "Deep Face Recognition" [PVZ15]. To not confuse it with the DeepFace model [Tai+14], which was published a short time before, it is commonly called "VGG-Face" model.

2.3.5 *COTS*

Two COTS FRSs have been deployed for verification purpose, which will be called COTS1 and COTS2. COTS FRSs are proprietary software, only returning verification scores and thresholds. Therefore, both COTS FRSs have only be used for the verification in the vulnerability analyses, not for image pre-selection. A typical automated border control scenario rather includes such a COTS FRS than an open source FRS. Therefore, the vulnerability of a morph within a COTS system is of high relevance for security considerations regarding morphing attacks. The name of the COTS FRSs will not be disclosed.

2.4 SOFT BIOMETRICS MODELS

Similar to embeddings of FRSs – i.e., embeddings to verify identities – the study at hand also aimed at analyzing embeddings of models which were trained to predict soft biometrics. For this aim, the LightFace [SO20] implementations of age, gender and race models were deployed. Those were retrained models (i.e., transfer learning) of the VGG-Face model (Section 2.3.2, Fig. 4). All convolutional and max-pooling layers were frozen for transfer learning, whereas the last four layers remained trainable. Furthermore, the last two layers were replaced by a fully connected layer and a softmax layer, of size 101 (age model), 2 (gender model), or 6 (race model), respectively.

The concept of the activation of the last layer (i.e., embeddings) is different for face recognition models than for models which predict a soft biometrics feature such as gender. Furthermore, the number of neurons in the respective last layers varied between age, gender, and race models. Therefore activations from an earlier layer of the respective models were extracted to retrieve embeddings for the respective soft biometrics. For this, the third last layer with 4,096 neurons each was chosen (Fig. 4), as this was the last layer until the coarse structure of the models (but not the weights and therefore not the activations) were identical.

2.5 DISTANCE METRICS

Different metrics were applied to calculate the distances between face embeddings for image pre-selection and identity verification. For pre-selection, the first sample of each data subject was used, and the distances between all data subjects were calculated pair-wise. The pair-wise distance (or similarity) values were aggregated into a distance (or similarity) matrix (Fig. 5). In track I, three different distance (or similarity) metrics were used for pre-selection (Sections 2.5.1–2.5.3). In track II, only Cosine distance was used for pre-selection (Section 2.5.2). In all tracks, Cosine distance was deployed for verification (Section 2.5.2) of the open source FRSs. The distance matrix in the case of image pre-selection was then used to find the closest subjects in embedding space (see Section 2.6 for more details).

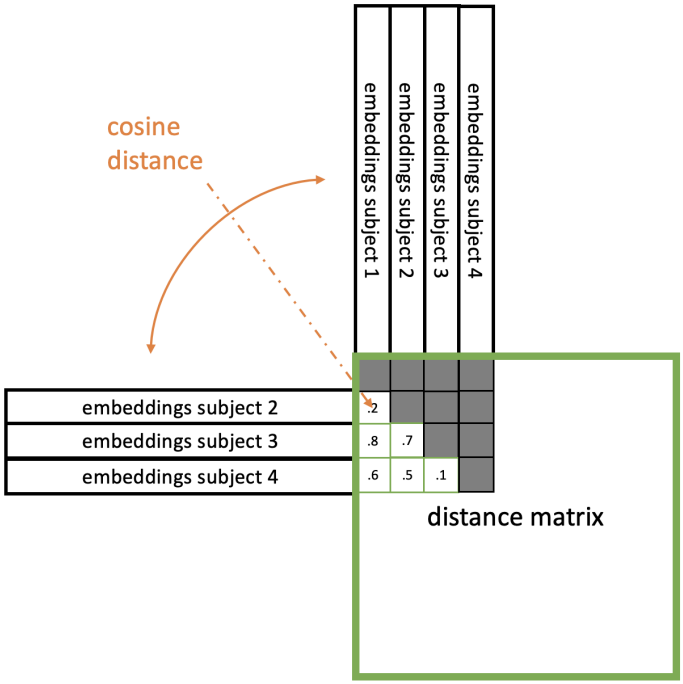

Figure 5: Scheme of distance calculations. The embeddings vectors of two non-mated data subjects were used to calculate a distance (e.g., Cosine distance). The resulting distance was saved into the distance matrix. Pair selection has been performed based on these distances.

2.5.1 *Euclidean distance*

The Euclidean distance [Mal13] between embeddings of two images A and B is the Pythagorean theorem in many dimensions, e.g., 512 dimensions for the ArcFace embeddings (Tab. 2).

$$d_{eucl}(A,B) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \tag{1}$$

in which

a_i i-th element of the embedding vector of image A

b_i i-th element of the embedding vector of image B

n length of the embedding vector

The Euclidean distance is defined in range $[0;\infty]$. Sometimes, the L2-norm or the dot product on the normalized embeddings vectors is used [Tai+14].

2.5.2 *Cosine distance*

The Cosine similarity [NB10] is defined by

$$s_{cos}(A, B) = \frac{a^T b}{||a|| \cdot ||b||} \tag{2}$$

in which

a embedding vector of image A

b embedding vector of image B

and is defined in range $[-1; 1]$. Correspondingly the Cosine distance is

$$d_{cos}(A, B) = 1 - s_{cos}(A, B) \tag{3}$$

2.5.3 *Mutual Information Score*

Mutual Information Score (MIScore) is a measure of similarity between two discretized vectors [VEB09]. Here, face embeddings have been discretized into 10 bins, and Mutual Information Score has been calculated between each pair of face embeddings. Mutual Information Score is symmetric, therefore exchanging A and B returns identical distances, making it suitable as similarity metrics.

$$MI(A, B) = \sum_{i=1}^{|a|} \sum_{j=1}^{|b|} \frac{\left|a_i \cap b_j\right|}{n} \log \frac{n \left|a_i \cap b_j\right|}{\left|a_i\right| \left|b_j\right|} \tag{4}$$

MI Mutual Information Score

$|a|$ length of embedding vector of image A

$|b|$ length of embedding vector of image B

2.6 PRE-SELECTION ALGORITHM

First, distance matrices were calculated based on a particular distance metrics (e.g., Cosine distance). Only one triangle of the matrix was selected. Then, the lowest off-diagonal value was selected to identify the face pair with the lowest distance. Then, it was checked if the participating data subjects match according to their soft biometrics labels, i.e., if they have the same gender, ethnicity (only in the UNCW data set), and if their age does not differ more than 5 years (only in the UNCW data set). If soft biometrics matched, the images were morphed. Both data subjects were removed from

the distance matrix as a result (i.e., by removing row and column of these data subjects). If soft biometrics did not match, only the element under evaluation was deleted from the distance matrix, therefore both data subjects were still available to be matched with different data subjects. This method was continued until all data subjects were either matched, or did not find a corresponding data subject within the soft biometrics prerequisites.

2.6.1 *Alternative: K-means constrained clustering*

In track I, an alternative algorithm to create pairs has also been tested. The algorithm was a deviation from a k-means clustering method, in which the number of samples in a leaf can be constrained (K-means constrained) [BBDoo]. The similarity of two samples was evaluated using the Euclidean distance between samples. The number of samples in a leaf was set to $1 - 2$, therefore, in all leaves with 2 samples, both data subjects in the leaf have been morphed.

2.7 MORPHING ALGORITHMS

In the study at hand, three out of four morphing algorithms were landmark-based (Alyssaq, NTNU, & UBO morpher), and had the following commonalities, or were used according to the following fashion:

- Morphing was based on averaging the two morph candidate images' positions of facial landmarks. 68 facial landmarks were extracted by the OpenCV dlib library [KS14], using an ensemble of regression trees to estimate their positions. Figure 6 shows the landmarks projected to an exemplary face image.
- A morphing factor (alpha) of 0.5 was applied, returning an "average" face, rather than shifting the landmark position of the morphed image disproportionately to one of the single morph candidates.
- No image preprocessing or postprocessing has been done other the steps included in the respective morphing packages. Rescaling and cropping was however performed for face recognition.

Exemplary morphs for all used morphing algorithms are displayed in Figure 7.

2.7.1 *Alyssaq morpher*

Alyssaq morpher (*FaceMorpher*, version 1.0) is an open source python implementation by Alyssa Quek to morph two or more faces [Que]. The created morphed face is returned on a black background.

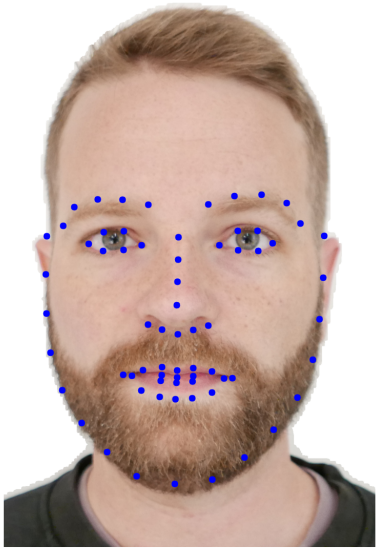

Figure 6: Dlib facial landmarks (blue dots) [KS14] on an example face image. 68 landmarks were estimated.

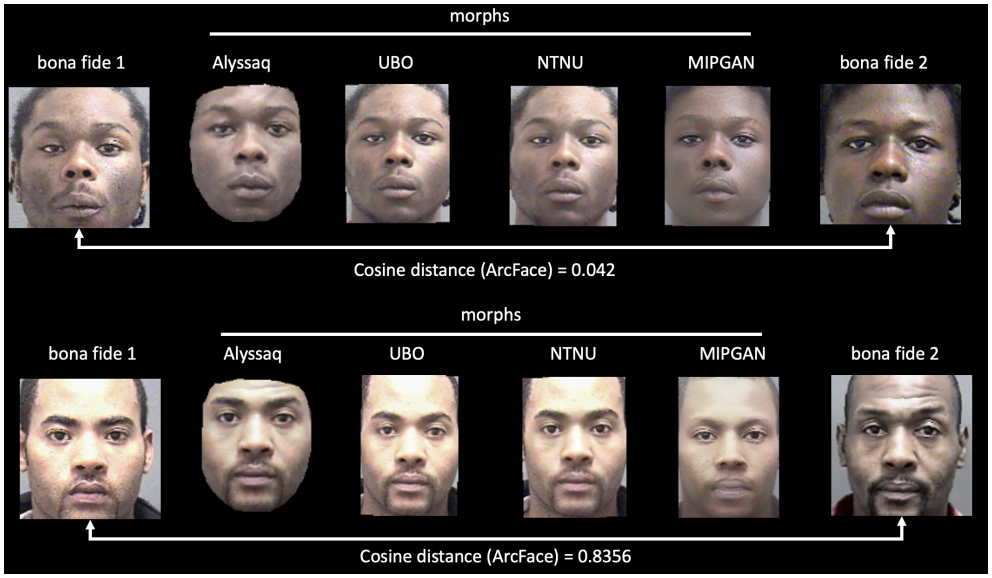

Figure 7: Exemplary morphed face images. The very left and right images represent two bona fide face images, respectively. Those face images were morphed using different morphing algorithms (central columns). The upper row represents images of two subjects, which had a low prior distance (i.e., high similarity of the bona fides prior to morphing) of the bona fide face images. The lower row represents images of two subjects with relatively high prior distance, i.e., low similarity before morphing.

2.7.2 *UBO morpher*

UBO morpher is a landmark-based morphing algorithm developed at University of Bologna (UBO) [FFM14; FFM16; FFM18; FFM19]. It was designed to investigate morphing attacks and automatically generates high quality morphed images. The package is not publicly available.

In addition to the 68 dlib landmarks (Fig. 6), UBO morpher required the position of centers of the left and right eye. For this, the 6 landmarks around each of the respective eye were used. The respective center was determined by the arithmetic mean of all x- and y-coordinates of each eyes' landmarks. The images were further retouched and color equalized. The created morphed face images were projected back to the image background of one of the two morph candidates.

2.7.3 *NTNU morpher*

NTNU morpher was developped at Norwegian University of Science and Technology (NTNU), and is – similarly to the UBO morpher, not publicly available. It conducts morphing based on facial landmarks of two morph candidates [Rag+17; Raj+20].

NTNU morpher processed the images as follows: First, the images were cropped to 1400×2100 portrait. Then 68 landmarks were estimated using dlib model [KS14]. Each respective two images were then morphed by calculating the average landmark positions, and calculating Delaunay triangles for both images. Both images were transformed to the average landmarks with an affine transformation of the triangles. Hue and saturation were copied from the first to the second image. Lightness was adjusted by smoothing histograms of both images, and by then adding the difference of the maximum location in both histograms to one of the two images. The average (morphed) face was projected on one of the backgrounds of the original images. The eyes were replaced to avoid double-iris artifacts. The images were further cropped. OpenCV Seamless Cloning [3] [Bra00] was performed to project the morphed images back to the background of both original face images, respectively. Therefore a total of two images were returned per morph pair.

2.7.4 *MIPGAN*

Lastly, one alternative morphing algorithm was deployed, which did not rely on dlib facial landmarks, but was build upon GANs.

Morphing through Identity Prior driven Generative Adversarial Network (MIPGAN) [Zha+21] is derived from StyleGAN [KLA19]. MIPGAN directly morphs two face images in the latent space, and generates a new face im-

3 Without the OpenCV Seamless Cloning, the morphed images comprised a lot of color artifacts.

age based on the morphed embeddings [Zha+21]. Here, MIPGAN2 [Zha+21] was used[4], which is based on StyleGAN2 [Kar+20].

2.8 CALIBRATION OF THE DECISION THRESHOLDS

To obtain comparable performance evaluations for different biometric systems, a common metric had to be deployed. Biometric systems are often trained or used in different data sets. Therefore, a particular verification threshold τ on the similarity or distance score can be determined individually for a data set. The score might be dependent on the particular use case. For instance, in a use case in which the number of false positive verification (i.e., False Match Rate (FMR)) must be very low, the proportion of false negatives (i.e., False Non-Match Rate (FNMR)) might increase likewise. FRONTEX [Fro15] proposed a FMR of $\leq 0.1\%$ for secure biometric systems.
For the FRGCv2 and the UNCW data sets, as well as for each individual open source FRS for verification, the decision thresholds were defined seperately and as follows: First – and only in the UNCW data set – a subset of 500 data subjects were sampled. From those subjects (respectively all subjects in FRGCv2 data set), all possible one-to-one combinations of mated pairs were compared using the respective FRS. Then, all possible combinations of non-mated comparison scores were calculated. Because the total amount of possible non-mated comparison highly outnumbered the amount of possible mated comparison scores, a uniform sampling from all possible non-mated pairs was performed. In the end, the number of mated comparisons equaled the number of non-mated comparisons.

Detection Error Trade-off (DET) curves were calculated for each FRS. According to ISO/IEC 19795-1:2021(E) [ISO21] DET curves visualize the relationship between false-negative and false-positive errors of a binary classification system as the discrimination threshold varies. From the respective DET curves, the decision thresholds τ for FMRs of 0.1% were empirically determined for each FRS (see Fig. S1 & Fig. S2) using PyEER [Mar]. All thresholds, alongside with the corresponding FNMRs are illustrated in Table 3.

For the COTS FRSs, a default threshold for a FMR of 0.1% was used, and not separately determined as with the open source FRSs.

2.9 VERIFICATION SYSTEM PERFORMANCE METRICS

The resulting morphed images were verified against bona fide probe images of the two contributing data subjects. For images of track I, track II, and track III, prodAvgMMPMRs (eq. 6) were calculated. In addition, for images of track II, Relative Morph Match Rate (RMMR) (eq. 7) and Morph Vulnerability Rate (MVR) (eq. 6) were calculated.

The following metrics were defined by [Sch+16; Sch+17; ISO21]. In the present study, all rates will be reported as decimal fractions, therefore distributing in the interval $[0; 1]$.

4 MIPGAN1 was also tested, but the resulting images exhibited an abundance of artifacts.

data set	FRS	FMR 0.1% (≙0.001) threshold	FNMR at FMR 0.1% (≙0.001)
FRGCv2	ArcFace	0.4765	0.00326
UNCW	ArcFace	0.4982	0.05105
	DeepFace	0.1245	0.784
	VGG-Face	0.146	0.318
	MagFace	0.666	0.0035

Table 3: The verification thresholds on the unnormalized distances for each open source FRS, calculated with the FRGCv2 and the UNCW data sets. The corresponding FNMRs are illustrated next to the thresholds as decimal fractions. The thresholds used to calculate product Average Mated Morph Presentation Match Rate (prodAvgMMPMR) have been set at a FMR of 0.1% (≙0.001 decimal fraction). The corresponding DET curves are illustrated in Figures S1 and S2.

2.9.1 *MMPMR*

The Mated Morph Presentation Match Rate (MMPMR) [Sch+16] is defined for similarity scores (eq. 5). The appropriate formulation for the distance scores replaces the $>$ sign with a $<$ sign.

$$MMPMR = \frac{1}{M}\sum_{m=1}^{M}\{(\min_{n=1,\ldots,N_m} S_m^n) > \tau\} \tag{5}$$

in which

M total number of morphed images

S_m^n mated morph comparison score of subject n at morph m

N_m total number of subjects constituting to morph m

τ decision threshold

2.9.2 *prodAvgMMPMR*

The prodAvgMMPMR [Sch+17] is a version with allows for a more probabilistic interpretation about the success of morph attacks

$$prodAvgMMPMR = \frac{1}{M}\sum_{m=1}^{M}[\prod_{n=1}^{N_m}(\frac{1}{I_m^n}\cdot\sum_{i=1}^{I_m^n}\{S_m^{n,i} > \tau\})] \tag{6}$$

in which, additionally to above,

I_m^n number of samples of subject n within morph m

$S_m^{n,i}$ mated morph comparison score of sample i of subject n at morph m

An example: One morphed image was evaluated. Two data subjects contributed to the morph with one image each. Three bona fide samples per subject were tested against the morph. For one data subject, $\frac{2}{3}$ of the comparison scores exceeded the threshold τ. For the other data subject, $\frac{3}{3}$ comparison scores exceeded the threshold τ. The prodAvgMMPMR then was simply the product of $\frac{2}{3}$ and $\frac{3}{3}$, therefore $\frac{2}{3}$.

2.9.3 *RMMR*

The RMMR metric [Sch+17] on the other hand takes the FNMR of a biometric system into account. Different biometric systems, calibrated at a particular FMR, can have different FNMRs. If the FNMR is high, the system is less suited for an operation in a particular scenario, e.g., access control. Consequently, it might produce low MMPMR or prodAvgMMPMR – therefore be less vulnerable to morph attacks – but at the same time rejects a large proportion of mated verification attempts. Therefore the RMMR relates the MMPMR to the FNMR.

$$\begin{aligned} RMMR &= 1 + MMPMR - TMR \\ RMMR &= 1 + MMPMR - (1 - FNMR) \\ RMMR &= MMPMR + FNMR \end{aligned} \tag{7}$$

FNMR, True Match Rate (TMR), and MMPMR are specific for the chosen decision threshold τ. If MMPMR is high, therefore if the morphs fool the FRS at τ, and at the same time if the FRS performs well by having a low FNMR, the RMMR would be around 1. On the other hand, if both the attack quality is low (low MMPMR), and the FRS also performs weakly by having a high FNMR, the RMMR would still be around 1. On the other hand, it the quality of the attack is poor (i.e., low MMPMR), and the FRS performs well by having a low FNMR, the RMMR would be around 0. For completeness, if the attack is of high quality (high MMPMR), and the FRS performs poorly (high FNMR), the RMMR could theoretically go up to 2. However, that would require the morphed comparison distances to be lower than the mated comparison distances.

2.9.4 *MVR*

Furthermore, a MVR represents a matrix-like representation of the success of a data set D of morphed images, evaluated on different FRS [Fer+22][5]. All FRSs (e.g., 6 different systems) verify the same number of bona fide images (e.g., 4) of each subject against the respective morph. $MVR_{4,6}^{D}$ then represents

5 The MVR metric was first used within the image Manipulation Attack Resolving Solutions (iMARS) project (https://cordis.europa.eu/project/id/883356).

the $4x6$ matrix where element (i, j) indicates the decimal fraction of morphed images, for which at least i verification attempts have been successful by both contributing data subjects, and for at least j FRSs (Fig. 8).

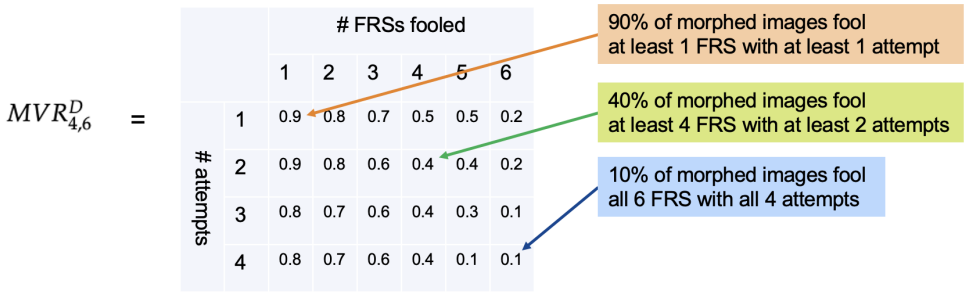

Figure 8: Morph Vulnerability Rate (MVR). The MVR is a matrix describing the success of a data set of morphed images. Several FRS (x axis) are attacked with several mated morph attack attempts (y axis). The element of a MVR matrix describes the proportion of successful verifications of both attackers (i.e., both contributing subjects of each contributing morph) at a given number of attempts (i.e., number of different bona fide images for both subjects) and with a particular number of fooled FRSs. Note that MVR was calculated as decimal fraction within range $[0; 1]$.

2.10 MULTILEVEL MODELING

In order to gain some insights from the morphed images, multilevel modeling has been performed on the morphed images of track II (Section 2.1.2). Only randomly pre-selected morphs have been used. The general intuition was to describe the prodAvgMMPMR as a function of several predictor variables. prodAvgMMPMR was used because it allows for a probabilistic interpretation of the morph vulnerabilities (eq. 6). However, a substantial degree of data points in such an analysis would be co-dependent, that is, they were generated by the same data subjects. Therefore, in addition to usual (fixed effects) predictor terms, a random (effects) intercept has been added for each data subject. Modeling was performed using R (version 4.0.3 (2020-10-10)) [To13] packages lmerTest (version 3.1-3) and lme4 (version 1.1-27.1) [LB90; Bat+15].

The following model was specified:

$$prodAvgMMPMR_i = \beta_0 + \beta_1 \cdot \alpha_i + \beta_2 \cdot \gamma_i + \beta_3 \cdot \rho_i + \beta_4 \cdot \mu_i + \beta_5 \cdot \psi_i + u_{i,1} + u_{i,2} + \epsilon \tag{8}$$

β_0	(fixed effect) intercept
β_k	(fixed effect) slopes
i	index of morphed image
$prodAvgMMPMR_i$	prodAvgMMPMR of morph i
α_i	average age of both subjects contributing to morph i
γ_i	gender of the subjects contributing to morph i
ρ_i	race of the subjects contributing to morph i
μ_i	morphing algorithm used
ψ_i	FRS used for verification
$u_{i,j}$	(random effects) intercept of subject j of morph i
ϵ	random error

Whereas all β_k were fixed effects regression coefficients, $u_{i,j}$ followed a Normal distribution centered around 0 with variance σ_u, and ϵ_i followed a Normal distribution centered around 0 with variance σ_ϵ. Consequently, any data subject j received their own (random) intercept, indicating a – in general – higher prodAvgMMPMR (if positive) or lower prodAvgMMPMR (if negative) than the other data subjects. Parameter estimation was performed minimizing log-likelihood rather than using REML criterion. Regarding ψ_i, only prodAvgMMPMRs were used which have been calculated using ArcFace and MagFace.

2.11 MORPHING ATTACK DETECTION

MAD can be conducted in different ways. Single-image Morphing Attack Detection (S-MAD) approaches evaluate the nature (i.e., bona fide or morphed) of an image by classification approaches of the presented face image [Sch+20]. On the other hand, D-MAD approaches compare a presented image with a trusted bona fide capture to evaluate the nature of the presented image [Sch+20]. In the present study, a D-MAD has been performed with the resulting suspected images. The presently used D-MAD approach was introduced by Scherhag et al. [Sch+20] and used differential analysis of ArcFace embeddings to train a Support Vector Machine (SVM) classifier. More specifically, the ArcFace embeddings have been extracted for suspected images (to be analyzed), and for of bona fide probe images of one of the participating morph candidates (Fig. 9). Both vectors were subtracted from each other. The resulting difference vectors of length 512 portray the samples of morphed (differential) images. As samples of bone fide (differential) images, the same procedure has been done subtracting the embeddings of two different bona fide captures of the same data subject. The resulting difference vectors have been scaled to follow a standard Normal distribution with $\mu = 0$ and $\sigma = 1$. In the present study, this procedure has been closely following [Sch+20] using ArcFace embeddings. Further the same procedure has been repeated in an analogous way using MagFace embeddings (Fig. 9).

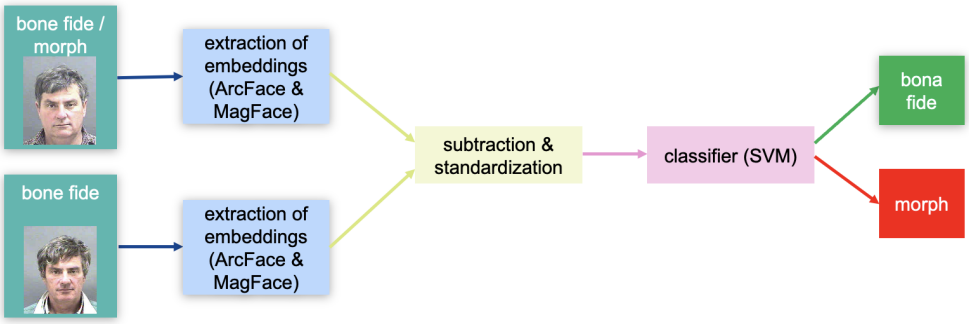

Figure 9: D-MAD pipeline. ArcFace or MagFace embeddings were extracted from bona fide images and morphed images. Differential embeddings have been created by subtraction of either the embeddings of a bona fide image from a morphed image or by the subtraction of a bona fide image from a different bona fide image. The differential vectors have been re-scaled to $N(0,1)$. A classifier was trained (on ArcFace and MagFace differential embeddings, separately) to differentiate between bona fide images and morphed images.

Only morphs of *track II* have been used for MAD training and testing. According to [Sch+20], a strict separation of morphed samples from the originating subjects has to be conducted, to avoid overfitting of the trained classifier. Therefore, for the bona fide distances, the subjects have been used for calculation which have not been used for morphing (i.e., because they did not meet the criteria such as having not more than 5 bona fide samples in the raw data set). For training, only a subset of 80% of randomly pre-selected morphs were used. i.e., without pre-selection based on any embeddings similarities. Further, morphed images of all morphing algorithms have been used together. Therefore, of a pair that has been morphed, the morphs of all four used morphing algorithms were put either in the training or testing set. A subset of 80% of all non-morphed data subjects (having at least 2 face samples, which were around 10,000 data subjects) have been used for training on the bona fide differential embeddings, and accordingly 20% for testing. For testing, further the remaining subset of the randomly pre-selected morphs have been used, and separately all morphs which were created based on pre-selection from embeddings from different FRS, such as ArcFace, DeepFace, VGG-Face, & MagFace.

A subject was either only in the training set, or in the testing set, when random pre-selection is conducted. However, because the same subjects have been used for different pairings, when different pre-selection was conducted, a training subject also appeared in the testing sets (of non-randomly pre-selected morphs), but as part of a different morph. Further, because the difference vectors between morphs and bona fides were calculated, the data should be less dependent than if the raw embeddings would have been used. This bias was unavoidable in the present analysis pipeline. However, if a bias is still present, the performance (i.e., accuracy) of a trained machine might be over-fitted on the testing sets with non-random pre-selection. Therefore, non-randomly pre-selected morphs should be slightly easier recognized by

this machine as a morphed image than they would be in independent data sets

2.11.1 *Testing metrics regarding morphing attack detection*

To evaluate MAD algorithms, ISO/IEC 30107-3 [ISO17] proposes to calculate Attack Presentation Classification Error Rate (APCER) and Bona fide Presentation Classification Error Rate (BPCER), which will be illustrated in the following. Similar to the metrics in Section 2.9, all rates will be reported in a range of $[0; 1]$.

- **APCER**: proportion of attack presentations using the same presentation attack instrument species incorrectly classified as bone fide presentations in a specific scenario
- **BPCER**: proportion of bona fide presentations incorrectly classified as presentation attacks in a specific scenario

APCER subserves as security measure, i.e., the proportion of attack presentation incorrectly classified as bona fide presentations should be low for a secure biometric system. In the contrary, BPCER subserves as convenience measure, i.e., a low number of false negatives is wanted in a biometric system which is operational.

Oftentimes, the BPCER at which the APCER is 10% ($= 0.1$) (BPCER10) is also reported [Sch+20]. BPCER10 can subserve as a convenience metric, at a given security level. For instance, if BPCER10 is 0.05, it means that at a given security level of detecting 90% of morphed images as such, 5% of bona fide images however get incorrectly classified as morphs.

2.12 FEATURE SELECTION ALGORITHMS

Feature selection was performed on the soft biometric embeddings vectors of track III. The aim was to evaluate the influence of each single element of the embedding vectors on prodAvgMMPMR, but similarly the influence of the gender, age, and race embeddings of the re-trained VGG-Face model separately (Section 2.4).

The feature selection algorithm was based on a design matrix $X \in \mathbb{R}^{n \times m}$ and a target vector $y \in \mathbb{R}^n$, whereas n is the number of observations, i.e., morphs, and m is the number of features, i.e. elements in a embeddings vector. The target vector y was chosen to be the prodAvgMMPMR of the morphs. Importantly, each row $X_{i,\cdot}$ of X was composed of the absolute difference between the embedding vectors of the two bona fide images which composed the morph:

$$X_{i,\cdot} = |e_{bf1} - e_{bf2}| \tag{9}$$

with e_{bf1} and e_{bf2} being the 4,096 dimensional (Tab. 2) embedding vectors (for each soft biometric model) of the two bona fide images, from which the morph was composed. The elements y_i of the target vector y were calculated as the prodAvgMMPMR of each morphed image, calculated by verifying bona fide images of both data subjects (eq. 6).

For this analysis, a large number of randomly pre-selected morphs have been created by Alyssaq morpher (Section 2.7.1), to generate a reasonable data dimension for feature selection. All data subjects have been split into either training or testing sets. ArcFace has been used for biometric verification.

For each soft biometrics (i.e., age, gender, & race) embeddings, all columns of the respective design matrices were discarded which had zero variance in the training set. This were around 50% to 85% of columns, depending on the particular soft biometrics model used. The remaining columns underwent a transformation to standard Normal distribution $N(0,1)$. The resulting design matrices underwent a feature selection using minimal Redundancy, Maximal Relevance (mRMR) (Section 2.12.1).

mRMR (Section 2.12.1) has been run for the design matrix composed of embedding differences for each model (i.e., age, gender, & race) individually. mRMR returned relative feature importances, i.e., the relative importance of each element of the embeddings vector, when included to the other, already selected, feature candidates. Using these relative feature importances, a rank of each feature (i.e., element of embeddings vector) has been created. Then, LGBM regression models (Section 2.12.2) have been constructed for feature sets of particular sizes. The sizes were of 20 equi-distant steps in $(0; N_{total}]$, with N_{total} being the number of maximally available features for the respective model after discarding features with zero variance. Each model was then trained on the respective feature set, and regression performance has been evaluated using Root Mean Squared Error (RMSE).

After having evaluated an optimal number of features from each kind of soft biometrics model, those respective features were concatenated into a huge design matrix to train another LGBM regression model (Section 2.12.2). The performance of this final, optimized model across soft biometrics types was then compared to the performances of the single models (i.e., using only age, gender, or race embeddings) without feature selection, and a combined model using the entire embeddings set of all soft biometrics models without feature selection. Chance prediction levels have further been calculated for each model by shuffling the predictions and calculating the respective RMSEs.

2.12.1 *mRMR*

An implementation of mutual information – often used in feature selection – is mRMR [DP05]. In mRMR, two terms are computed based on mutual information. First, the mutual information between a feature X_i and the outcome class c is computed. A high mutual information between these two variables indicates a high *relevance* of this particular feature, i.e., high predictability of

the target class c by the feature variable X. Second, the *redundancy* to other features is calculated. The redundancy term in this formulation is the average of mutual information of the feature to all other features, which have already been selected. Redundancy therefore is to be minimized, whereas relevance is to be maximized, when a new feature is selected. With the combination of those two terms, such as for example in equation 10, the relevance term of a feature is penalized by the redundancy to other, already selected, features.

$$mRMR(X_i) = I(X_i, c) - \frac{1}{|S|} \sum_{j \in S} I(X_i, X_j) \tag{10}$$

$$mRMR = \text{relevance} - \text{redundancy} \tag{11}$$

2.12.2 *LGBM*

LGBM is a gradient boosting algorithm based on decision trees [Ke+17]. Boosting generally combines several weak learning algorithms to combine them to a strong learner [Sch90]. It was developed among other things to allow for the analysis of larger data sets, and on the same hand, is supposed to return higher accuracies [Fri01].

3 RESULTS

3.1 EVALUATION OF DISTANCE METRICS FOR IMAGE PRE-SELECTION

First, different distance or similarity metrics were evaluated based on their ability to enhance the morph attack success rate of morphs created using a small data set (i.e., track I, Section 2.1.1). Pairs for morphing have been found based on different distance (or similarity) metrics, i.e., Euclidean distance, Cosine distance, mutual information (similarity), or with an entirely different pairing algorithm (k-means constrained). For comparison, pairs were generated randomly. Figure 10 illustrates the distributions of the prodAvgMMPMRs of the resulting morphs.

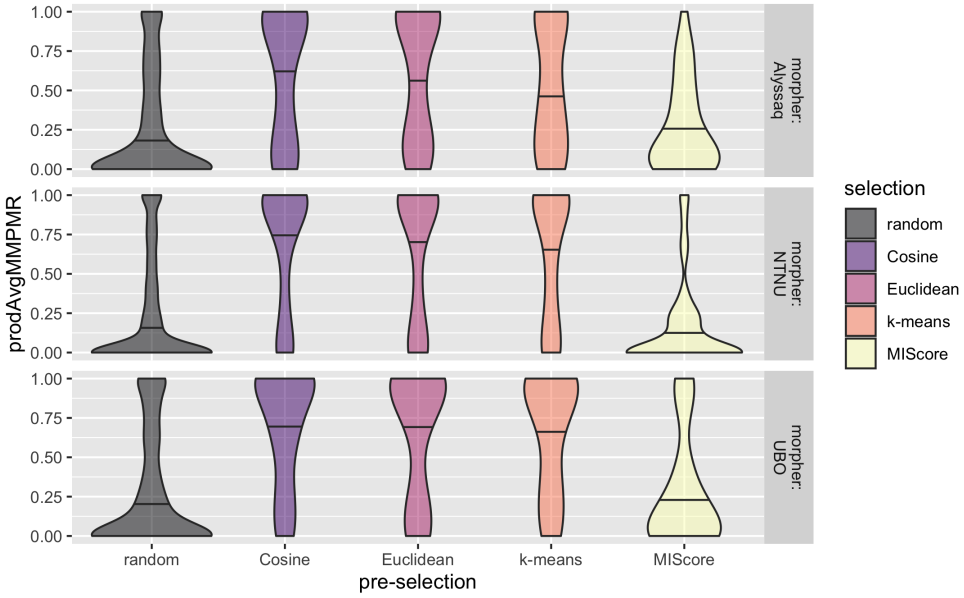

Figure 10: Mated morph comparison success rates for different image pre-selection metrics. prodAvgMMPMRs (y axes) are illustrated for different pre-selection methods (x axis & color-coded). Data density is plotted in horizontal direction. Median values are illustrated by horizontal black bars. The same pairs have been morphed by different morphing methods (rows). Random assignment of the morphing pairs are displayed in the left-most column. Cosine distance and Euclidean distance show higher median prodAvgMMPMRs, similar to images pre-selected by the k-means constrained algorithm. On the contrary, image pre-selection based on mutual information (MIScore) only slightly increased morph attack success rate. Note that prodAvgMMPMR was calculated as decimal fraction within range $[0; 1]$.

All pre-selection methods tested have improved morph attack success rate, compared to randomly generated morphs. However, Cosine distance and Eu-

clidean distance performed better than image pre-selection based on mutual information. Using a k-means constrained algorithm, morphs performed nearly as well as using Cosine and Euclidean distances. All tested morphing algorithms, Alyssaq, UBO, and NTNU morpher, profited from image pre-selection. However, morphs produced by UBO morpher and NTNU morpher showed higher median prodAvgMMPMRs than morphs produced by Alyssaq morpher. Notably, even morphs created by random face pairs produced positive values for prodAvgMMPMR.

3.2 EVALUATION OF FACE RECOGNITION SYSTEMS FOR IMAGE PRE-SELECTION

In a next series of analyses (track II, Section 2.1.2), embeddings of different FRSs have been evaluated based on their ability to increase morph attack success rate.

3.2.1 *Mated morph comparison performance*

Figure 11 illustrates the distribution of prodAvgMMPMR values, produced by face morphs which have been pre-selected using different FRSs, morphed with different morphing algorithms, and verified against bona fide probe images using again different open source FRSs.

In general, image pre-selection as compared to random pairing increases morph attack vulnerability when the resulting morphs are verified using ArcFace or MagFace. Using MagFace for verification, highest morph vulnerability was shown, followed by ArcFace. VGG-Face and more so DeepFace showed lowest vulnerability to morph attacks. See Section 3.2.2 for in detail analysis of this behavior.

In the following, the focus is narrowed down on morphs verified with ArcFace and MagFace: Figure 11 further illustrates, that image pre-selection based on embeddings from ArcFace and MagFace created best morphing attacks, followed by VGG-Face and lastly DeepFace. The exact ranking of ArcFace and MagFace for image pre-selection was dependent on whether ArcFace or MagFace has *also* been used for image verification. If the same FRS was used for verification, pre-selection based on the same system performed best. If COTS FRSs have been used for verification, this bias vanishes (Fig. S3). Further, using COTS FRSs, the prodAvgMMPMR is mostly accumulating around 1, indicating high vulnerability, even for morphs with random pre-selection applied (Fig. S3). However, the morphs created with MIPGAN and verified using COTS FRSs illustrate again the benefit of image pre-selection.

Importantly, the patterns described before were highly similar, regardless of which morphing algorithm has been used. However, there was a distinguishable difference in morph attack success. NTNU morpher and UBO morpher produced best morphing attacks, followed by Alyssaq morpher and lastly MIPGAN (Fig. 11 & S3).

The MVR has been introduced as a general measure of morph attack success across different verifying FRSs (Section 2.9.4). Briefly, the elements of

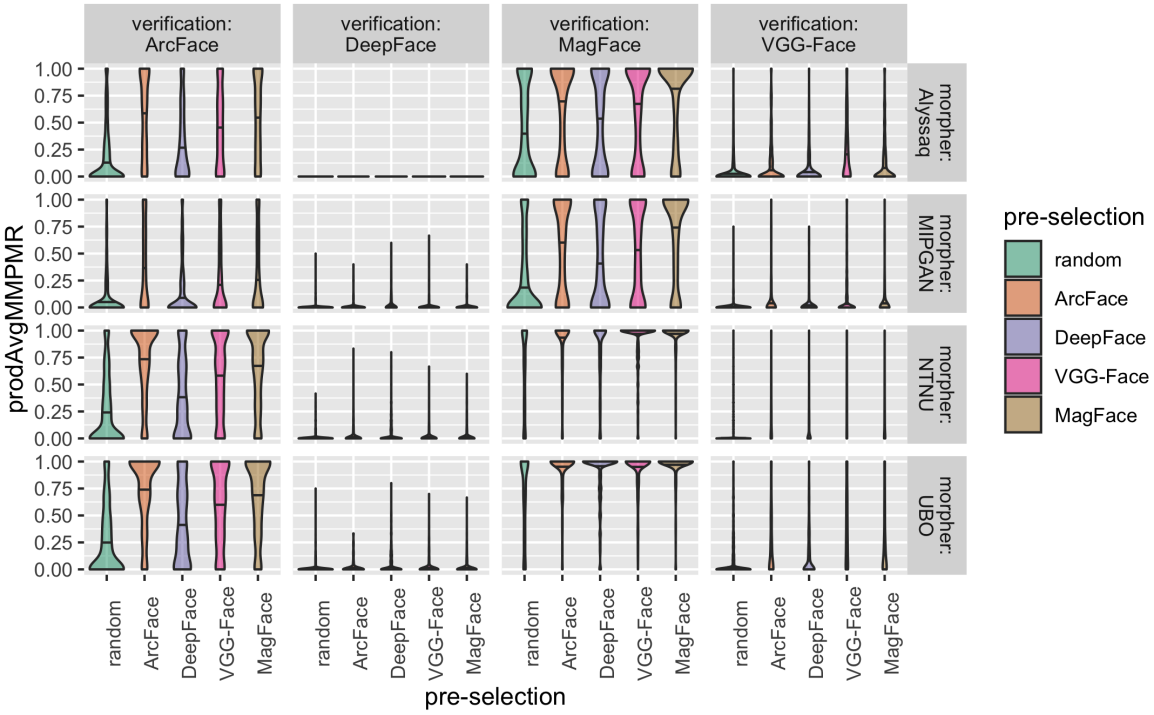

Figure 11: Mated morphs comparison success rates for different image pre-selection embeddings. prodAvgMMPMRs (y axes) are plotted for different pre-selection methods (x axis & color-coded). Density is plotted in horizontal direction. Median values are illustrated by horizontal black bars. The same pairs have been morphed by different morphing methods (rows). Random assignment of the morphing pairs are displayed in the left-most column. All morphs have been evaluated by different open source FRSs (columns). Note that prodAvgMMPMR was calculated as decimal fraction within range $[0; 1]$.

an MVR matrix contain the proportions of successful morph attacks (with both participating data subjects), at a particular number of FRSs and a given number of attempts. As higher the values, and as further high values spread towards the lower border and the rightmost border of the matrix, the more effective are morphs of a tested data set.

Figure 12 illustrates MVRs for morphs created by the UBO morpher. Again, using pre-selection generally increased the MVRs. All non-random pre-selection methods have lead to the successful outwitting of at least four (out of six) FRS in around 70 to 90% of cases with one attack attempt. Contrarily, random morphs only exceed 47%. In around 17% to 47% of cases, all four morph attack attempts were able to fool four different FRS, when pre-selection was conducted. However, only single-digit percentages of morphs were able to fool four FRSs with all four attack attempts.

The numbers were comparable when morphs were created by NTNU morpher instead of UBO morpher (Fig. S4). The numbers were considerably lower when morphs were created by Alyssaq morpher (Fig. S5), and even lower for morphs created by MIPGAN (Fig. S6). However, a definite dis-

tinction between morphs pre-selected by different FRS embeddings is more complex.

3.2.2 *Relative mated morph comparison performance*

To further examine the different pre-selection methods, as well as the behavior of the verification algorithms, the distributions of the raw distance scores of mated comparisons, non-mated comparisons, and mated morph comparisons (exemplary for the UBO morpher) have been visualized in Figure 13.

Across all four open source verification FRSs it can be seen, that the mated morph comparison scores were distributed between mated scores and non-mated scores. However, they were closer aligned to the mated scores than the non-mated scores, even for morph pairs without pre-selection (i.e., random assignment). However, all pre-selection methods were better than random pre-selection. Again, the same verification system showed preference for morphs pre-selected by its own embeddings before morphing.

However, the comparison decision highly varied between the verification FRSs. Whereas DeepFace falsely verified only a very small number of morphs successfully, followed by VGG-Face, ArcFace and most significantly MagFace falsely accepted nearly all morphs as mated comparisons. On the contrary, at the calibrated threshold of $FMR = 0.1\%$, DeepFace and to a less severe extend VGG-Face exhibited high FNMR, therefore falsely rejecting a large proportion of mated verification attempts (Tab. 3). However, ArcFace, and more strikingly MagFace, had very low FNMRs at the given FMR (Tab. 3). This has lead to a higher vulnerability of *better* – in terms of low FNMR at a given FMR – FRS for morphing attacks.

Figures S7, S8, & S9 show the respective distributions for morphs created with the other morphing algorithms. The general patterns were the same as in Figure 13. However, whereas the distance distributions of mated morph comparisons with NTNU morphs closely aligned to those morphs created with UBO morpher, both Alyssaq and MIPGAN showed higher distances, leading to higher amount of rejections of mated morphs at the given decision thresholds.

Whereas mated morph distances of *good* FRSs – such as ArcFace and MagFace – were distributed between mated distances and non-mated distances, Figure 13 indicates that there is a chance of separating morphs from mated comparisons by adjusting decision thresholds. This could dramatically lower the vulnerability for MagFace, where the distance distributions of mated and morphs showed only a slight overlap, and to a smaller degree for ArcFace, where there was still a strong overlap, and where threshold adjustment would lead to significant higher FNMR. Contrarily, the distributions of mated morph distances of *bad* FRSs such as DeepFace and VGG-Face closely aligned to the distribution of the mated distances (Fig. 13). In case of DeepFace, especially when image pre-selection as well as verification was performed with the same FRS, mated morph distances were even smaller than mated distances.

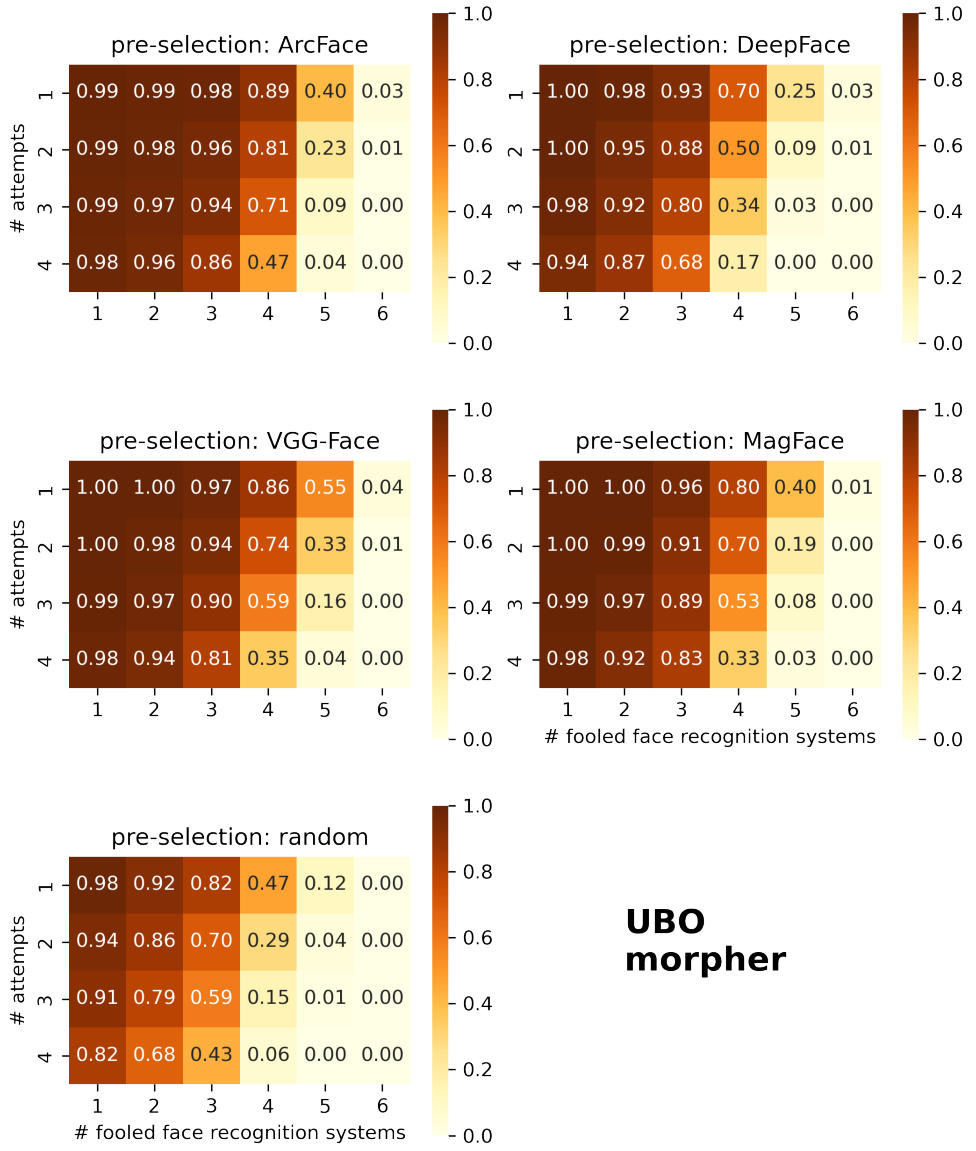

Figure 12: Morphs vulnerability rates (MVR) of morphs generated by the UBO morpher. Different FRSs have been used for image pre-selection, i.e., ArcFace, DeepFace, VGG-Face, or MagFace (different heatmaps). Alternatively, a random assignment of pairs has been conducted (bottom heatmap). For each FRS used for pre-selection, the resulting morphs have been verified against four bona fide images of each data subject. The ratio of successful attempts for both data subjects is illustrated on each y axis of each individual plot. Further, different FRSs have been used for mated morph verification, four open source FRSs and two COTS FRSs. The proportion of successful attacks across several FRSs is illustrated on each x axis. The MVR is indicated in each cell, as well as color-coded, and describes the proportion of successful verifications for a given number of attempts (y axes) and FRSs (x axes). Note that MVR was calculated as decimal fraction within range $[0; 1]$.

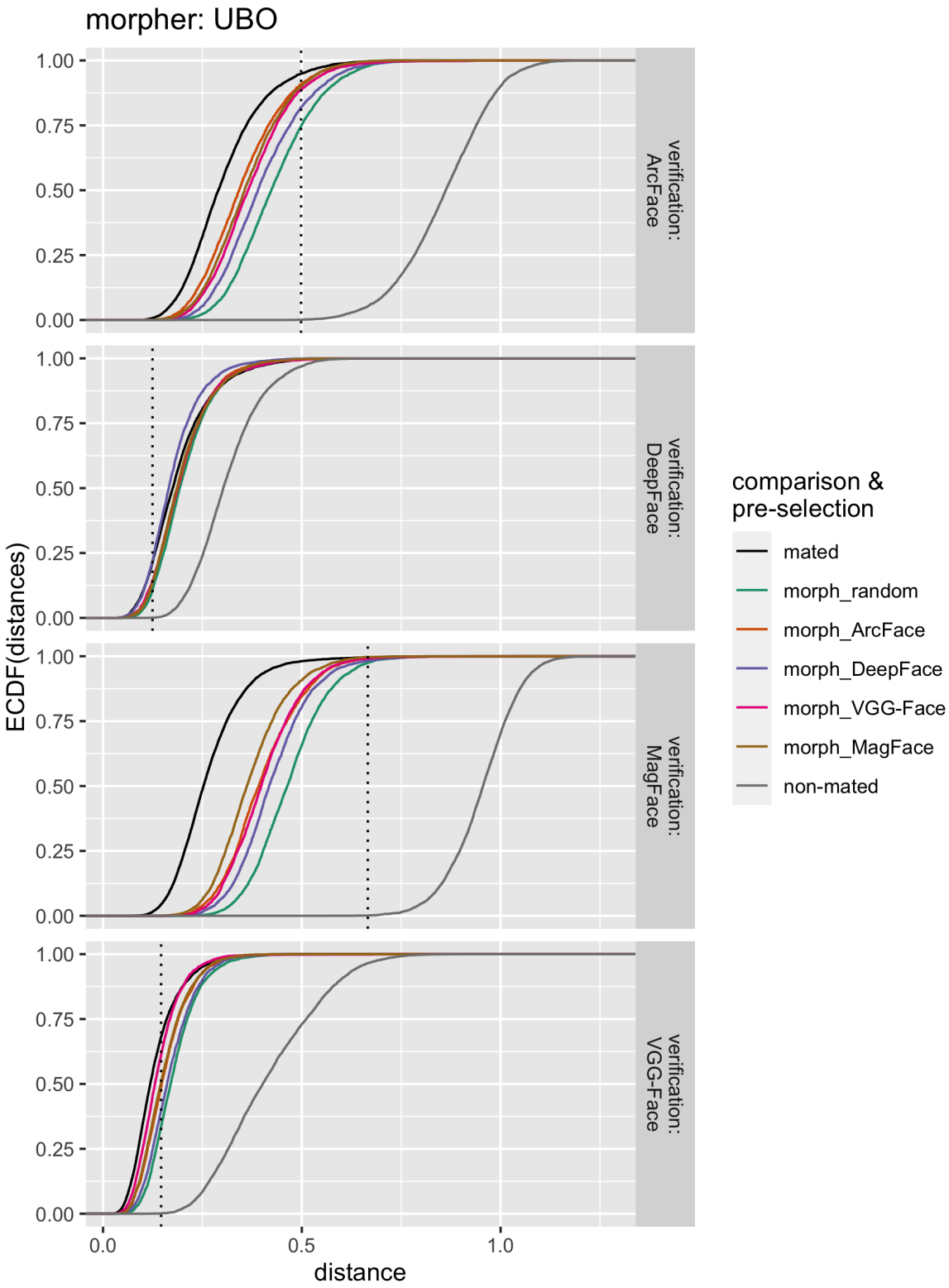

Figure 13: Empirical Cumulative Distribution Functions (ECDFs) for distance scores of the open source FRSs. Mated, non-mated, and mated morph comparisons have been conducted. Morphs were created by UBO morpher. The difference distances for the comparisons are displayed on the x axis. The (cumulative) proportion of positive verifications at a particular distance score is plotted at the y axes. Different FRS have been used for verification (rows). Color-coded are the different type of comparisons, i.e., mated or non-mated, or mated morph comparisons, with morphs pre-selected by the help of face embeddings of particular FRS. The dotted vertical lines indicate the 0.1% FMR threshold, individually for each FRS used for verification.

Figure 14 further illustrates the ECDFs of the similarity scores using COTS FRSs. As the COTS FRSs have not been used for pre-selection, the results are more *neutral* with respect to the pre-selection algorithm. First, all kind of morphs, even with random pre-selection, were likely to get positively verified by the COTS FRSs. However, similar to the open source FRSs, the distributions of the mated morph comparisons shifted towards the distributions of the mated comparisons, when pre-selection was applied. A hierarchy can be seen, between the different pre-selection methods. Morphs derived from a pre-selection with MagFace generated highest similarity scores, followed by ArcFace, VGG-Face, and lastly DeepFace.

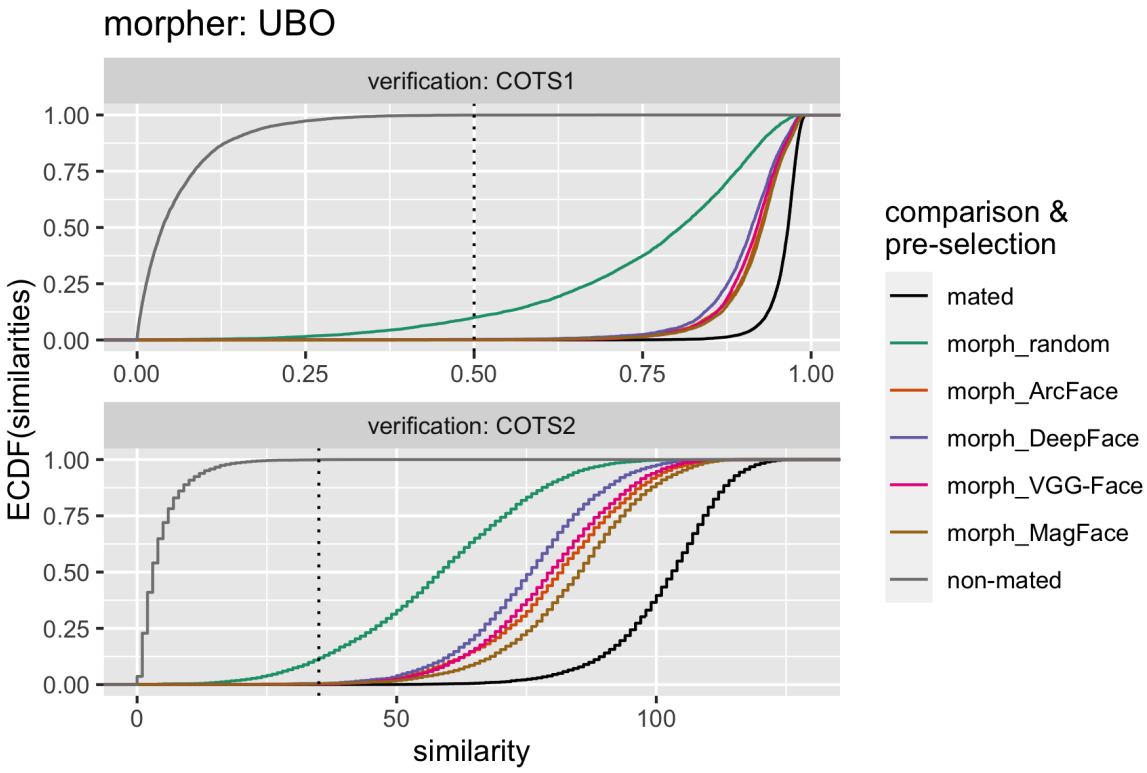

Figure 14: ECDFs for similarity scores of the COTS FRSs. Mated, non-mated, and mated morph comparisons have been conducted. Morphs were created by UBO morpher. The difference similarities for the comparisons are displayed on the x axis. The (cumulative) proportion of positive verifications at a particular similarity score is plotted at the y axes. Note that because similarities instead of distances are used, so the interpretation of the x axes must be flipped compared to Figure 13. Different COTS FRS have been used for verification (rows). Color-coded are the different type of comparisons, i.e., mated or non-mated, or mated morph comparisons, with morphs pre-selected by the help of face embeddings of particular FRS. The dotted vertical lines indicate the 0.1% FMR threshold, individually for each FRS used for verification.

To further take into account the performances of the single FRSs, the RMMR was calculated using the open source FRSs for verification. The RMMR corrects the MMPMR for the FNMR (eq. 7). Therefore, the strong inflation of mated

morph comparison values of the previous chapter can be corrected for, especially in FRSs with high FNMRs. Table 4 illustrates the RMMR values for differently pre-selected, morphed, and verified images. The pattern manifests, that if the same FRS is used for pre-selection and verification, the RMMR is highest in most cases. The second highest RMMR however often happened to derive from a pre-selection with MagFace, followed by ArcFace and VGG-Face. Higher RMMR can further be observed for morphs created by UBO morpher and NTNU morpher.

verification \ pre-selection	morpher	random	ArcFace	DeepFace	VGG-Face	MagFace
ArcFace	Alyssaq	0.24	0.64	0.39	0.55	**0.63**
DeepFace		0.78	0.78	0.78	0.78	0.78
VGG-Face		0.35	0.44	0.37	0.6	**0.45**
MagFace		0.44	0.64	0.52	**0.65**	0.76
ArcFace	UBO	0.32	0.79	0.51	0.65	**0.72**
DeepFace		0.79	**0.81**	0.84	0.8	**0.81**
VGG-Face		0.42	0.57	0.45	0.71	**0.58**
MagFace		0.72	**0.95**	0.87	**0.95**	0.97
ArcFace	NTNU	0.31	0.78	0.49	0.64	**0.72**
DeepFace		0.79	0.8	0.84	**0.81**	**0.81**
VGG-Face		0.4	0.53	0.45	0.71	**0.55**
MagFace		0.65	**0.91**	0.83	**0.91**	0.97
ArcFace	MIPGAN	0.13	0.44	0.22	0.32	**0.37**
DeepFace		0.79	0.79	0.8	0.79	**0.8**
VGG-Face		0.33	**0.37**	0.35	0.44	**0.37**
MagFace		0.28	**0.6**	0.45	0.54	0.68

Table 4: RMMRs. Images have been morphed using different morphing algorithms, pre-selected using embeddings of different FRSs or alternatively, been randomly pre-selected, and verfied using different FRSs. The RMMR corrects the MMPMR by the FNMR of the verification FRS (see eq. 7). The highest values row-wise have been highlighted in bold, leaving out the quasi-diagonal elements, i.e., if pre-selection and verification FRSs coincided. Note that RMMR was calculated as decimal fraction within range $[0; 1]$.

Table 4 can be summarized in the following fashion. To get some idea about how good the single pre-selection FRSs have performed across morphing algorithms and open source verification FRSs – using RMMR as a metric – each row of Table 4 has been converted to ranks (1 to 5). 5 indicated the FRS for pre-selection (columns) which had highest RMMR compared to the other elements, and 1 indicated the FRS with lowest RMMR, respectively. If same values occurred in a row, decimal numbers have been used. The ranks were then averaged across rows, therefore averaged across morphing algorithms and verification FRSs. Table 5 illustrates the average ranks for the different

pre-selection methods. Pairs based on MagFace embeddings had the highest chance of creating high RMMR values, followed by ArcFace, VGG-Face, and finally, DeepFace. Randomly pre-selected pairs perform worst across different morphing algorithms and verification systems.

pre-selection	average rank
random	1.1250
ArcFace	3.6250
DeepFace	2.6250
VGG-Face	3.5625
MagFace	4.0625

Table 5: Average ranks for RMMR values for the different pre-selection methods, i.e., random assignment or based on embeddings of 4 different FRSs.

3.2.3 *Morph attack detection performance*

A D-MAD algorithm – motivated by [Sch+20] – has been trained with randomly pre-selected morphs and tested with randomly and non-randomly pre-selected morphs. Differential embeddings have been calculated using ArcFace and MagFace embeddings in two seperate runs. Figure 15 illustrates the corresponding BPCER10 values of the MAD classifiers, tested on morphs with different pre-selection applied (and bona fide images). The operational point values BPCER10 were lower for the D-MAD classifier trained and evaluated with MagFace embeddings, than the ones trained and evaluated with ArcFace embeddings. Furthermore, BPCER10 on the test data sets were lowest for randomly pre-selected pairs for morphing, and higher if the test set contained morphs of pre-selected pairs. This trend was more pronounced in morphs created by NTNU morpher and even more in morphs created by UBO morpher. On the other hand, morphs created by Alyssaq morpher or MIPGAN did not lead to such a high rise in BPCER10 values.

Further, BPCER10 values were lowest for morphs with randomly pre-selected pairs, but higher if pre-selection was performed, especially if pre-selection was performed based on embeddings from MagFace or ArcFace. BPCER10 values were lowest when morphs of MIPGAN were tested, followed by morphs by Alyssaq morpher, and where highest for morphs created by UBO morpher and NTNU morpher. A high value of BPCER10 renders the MAD system inconvenient for practical purpose. The BPCER10 was increased by pre-selection (i.e., MagFace and ArcFace) and by the morphing algorithm used (i.e., UBO morpher and NTNU morpher). The trend is illustrated in more detail in Figure 16. Higher BPCER and APCER values were produced by the respective FRSs, if pre-selection has been performed, and especially if it has been performed using ArcFace or MagFace embeddings. This was consistent across different morphing algorithms.

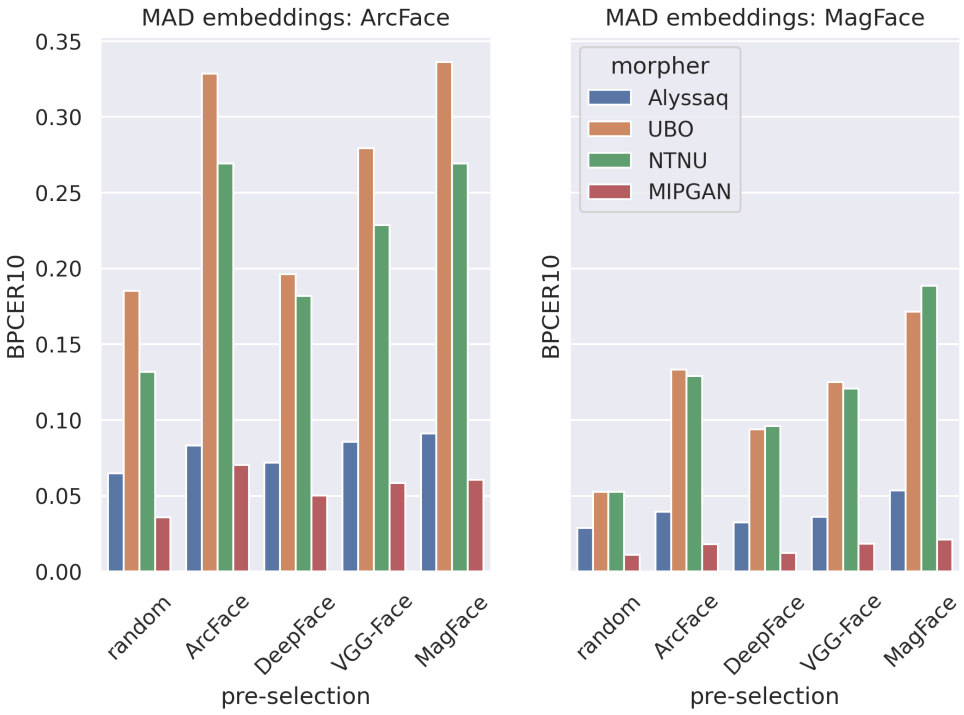

Figure 15: D-MAD algorithm performances. BPCER10 values of the classifiers tested on differently morphed and differently pre-selected testing data sets is shown. Left: metrics from an D-MAD algorithm trained with ArcFace embeddings. Right: metrics from an D-MAD algorithm trained with MagFace embeddings. The images morphed by different morphing algorithms are illustrated in different colors. The pre-selection method to generate the pairs for morphing are distributed along the x axes. Note that BPCER10 was calculated as decimal fraction within range $[0; 1]$.

High differences can be observed regarding which FRS is used to extract embeddings to train and test D-MAD classifiers. BPCER10 values were approximately half in size when MagFace was used for D-MAD, irrespective of which FRS was used to extract embeddings for image pre-selection (Fig. 15). On the other hand, the advantage of attacks morphed by UBO morpher over embeddings morphed by NTNU morpher disappears when MagFace is used for D-MAD compared to ArcFace (Fig. 15). The same can be seen in more detail in the DET curves (Fig. 16). The APCER and BPCER values were generally smaller, indicating a better performance of the MAD algorithm.

Interestingly, in some cases in Figure 16, it can be observed that there was not a consistent bias of the D-MAD algorithms towards being fooled from morphs pre-selected by embeddings of the same FRS than used for D-MAD. MagFace embeddings for pre-selection performed best in most cases to fool the D-MAD algorithm, even if it was trained with ArcFace embeddings.

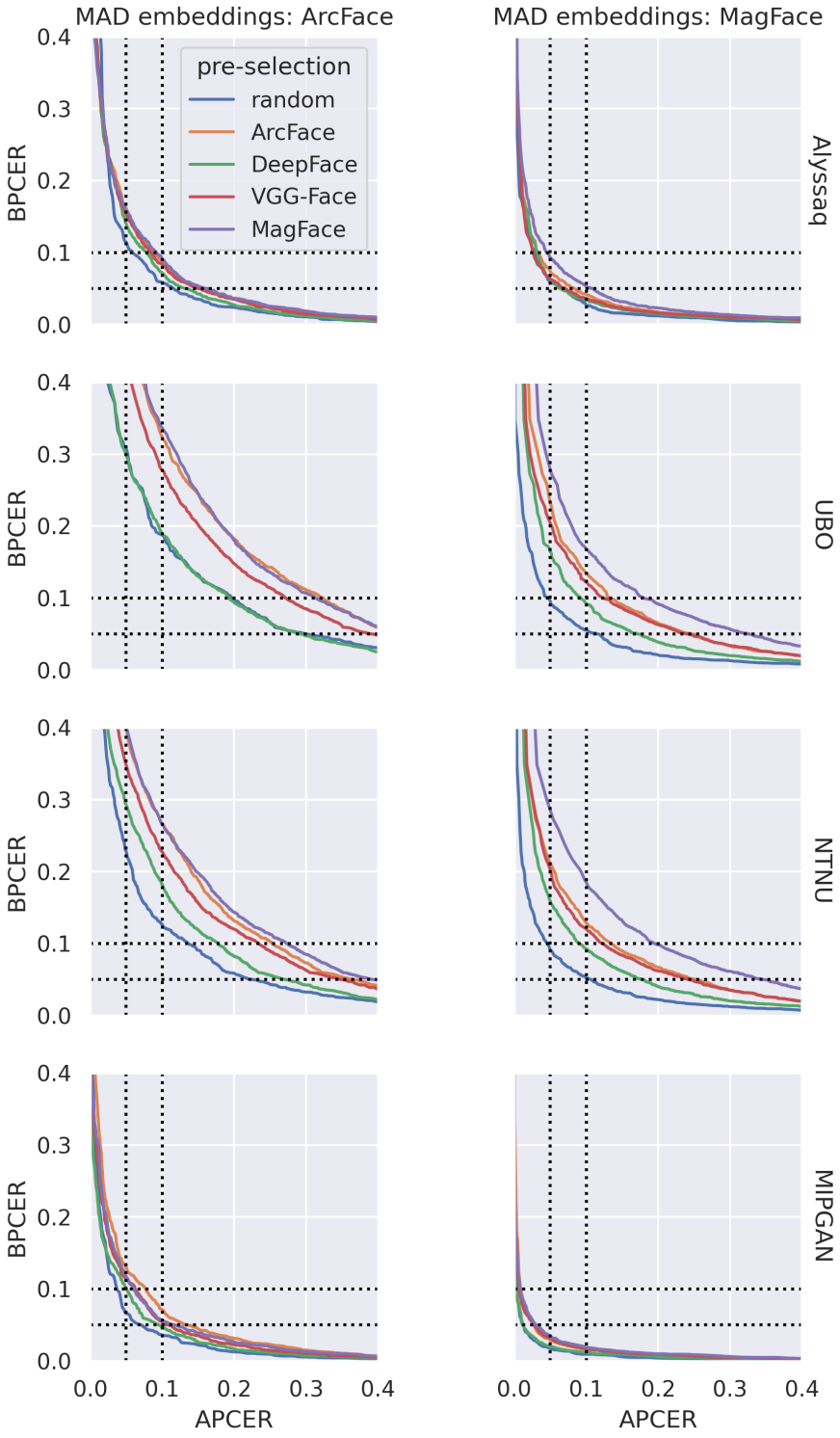

Figure 16: DET curves of the D-MAD approaches. Left column: D-MAD approach which deployed ArcFace embeddings (original version). Right column: D-MAD approach which deployed MagFace embeddings. Morphs of the different morphing algorithm are separated by rows. Data sub-sets of differently pre-selected morph pairs are color coded. The BPCER is plotted against the APCER. Dotted lines indicate the positions where BPCER or APCER are 0.1 (i.e., 10%) and 0.05 (i.e., 5%). Note that both rates were calculated as decimal fraction within range $[0; 1]$.

3.2.4 *Statistical analysis of morph attack success rates*

The prodAvgMMPMRs of randomly pre-selected morphs of the open source FRSs have been further analyzed using a multi-level linear model (eq. 8). Table 6 illustrates all fixed-effects predictor variables, alongside with their estimated magnitudes, t, and p values. The intercept represented a black, female data subject, aged 28 years, morphed with Alyssaq morpher, and the prodAvgMMPMR calculated using ArcFace. Only p values < 0.0001 will be interpreted, to account for a sufficient correction of multiple comparisons. The intercept was significantly positive, indicating an average prodAvgMMPMR of around 0.216 for this scenario. Using NTNU morpher, or in particular UBO morpher for morphing, increased the average prodAvgMMPMR by 0.134 or 0.174, respectively. Contrarily, using MIPGAN decreases prodAvgMMPMR by around 0.136. Interestingly, age, as linear predictor, had a negative impact on prodAvgMMPMR of -0.003. Therefore, 1 year of age increase leads to a reduction of prodAvgMMPMR by around 0.003. Accordingly, increasing age by 30 years would lead to an increase of prodAvgMMPMR by around 0.1, which is 10%. Further, gender and hispanic race did not change prodAvgMMPMR significantly. However, affiliation to caucasian race significantly decreased prodAvgMMPMR by 0.144, indicating less vulnerability of FRSs in this data set. Furthermore, using MagFace for verification significantly increased prodAvgMMPMR [1].

3.3 FEATURE SELECTION ON SOFT BIOMETRICS EMBEDDINGS

Feature selection and regression on embeddings of soft biometrics models to predict mated morph comparisons has been performed to evaluate their suitability for image pre-selection (track III, Section 2.1.3). Figure 17 illustrates the RMSE of regression models with differently sized feature subsets of the x most important features. For all types of soft biometrics models, the reduction of the numbers of features slightly reduced the testing error compared to a full model. However, the reduction was only minimal, as can be seen on the scaling of the y axis (Fig. 17). However, the number of features could effectively be reduced to a small fraction of the original number of features.

Furthermore, testing errors of resgression models with and without features selection are illustrated in Figure 18. Without feature selection, the errors of regressions based on age, gender, and race embeddings did not differ markedly, but there was a slight trend to lower errors for the race model. Combining the embeddings of all three models did not lead to a better prediction. Selecting an educated subset, i.e., a subset after feature selection did not improve the prediction, but rather lead to a subtle increase in the error.

1 The same analysis has also been repeated using the two verification FRSs separately, to account for possible interaction effects. The qualitative outcome of the analyses were identical to the results described here.

predictor	parameter estimate	t	p
intercept	0.216	6.1	< 0.0001
morpher_MIPGAN	-0.136	-13.7	< 0.0001
morpher_NTNU	0.134	14.5	< 0.0001
morpher_UBO	0.174	17.5	< 0.0001
age	-0.003	-4.0	< 0.0001
gender_male	-0.069	-2.0	0.05
race_hispanic	-0.153	-1.3	0.19
race_caucasian	-0.144	-4.0	< 0.0001
verification_MagFace	0.321	51.2	< 0.0001

Table 6: Multilevel model results. prodAvgMMPMR of randomly pre-selected morphs has been modeled as a function of several fixed effects predictor variables and subject-specific random intercepts (eq. 8). Fixed effects predictors are illustrated in columns, alongside their parameter estimates, t and p values. The model's intercept corresponds to a black, female data subject, aged 28 years, which was morphed with Alyssaq morpher, and the prodAvgMMPMR was calculated using ArcFace. Age was modeled as linear predictor in years. The other predictors were categorical, therefore either the morphing algorithm, gender, race, or verification FRS increased or decreased the prodAvgMMPMR on average by the value of the respective parameter estimate.

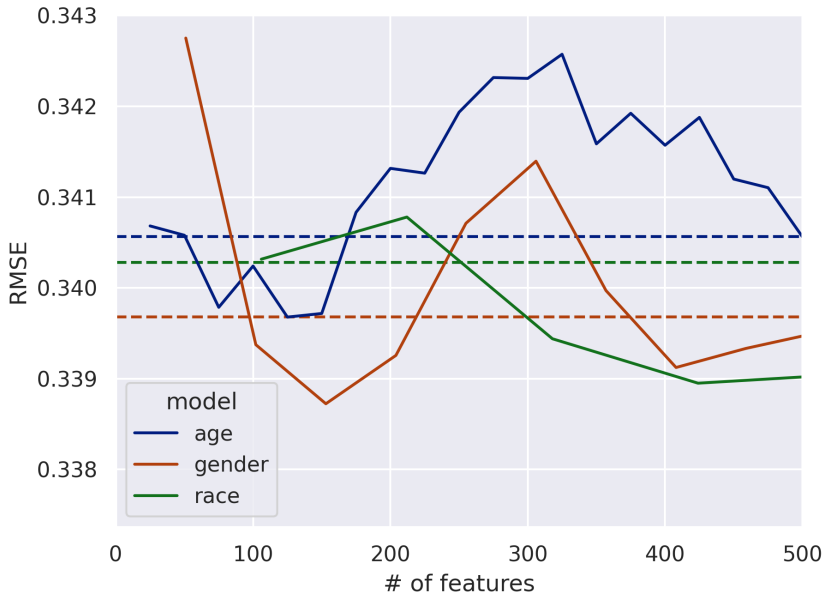

Figure 17: Testing errors for morph attack success prediction on different feature subsets of embeddings of the different soft biometrics models. The x axis indicates the number of (most important) features used for regression. Importance of the features has been evaluated using mRMR. The y axis corresponds to the RMSE on predicting prodAvgMMPMR. The soft biometrics models to extract embeddings for prediction are separated by color. Dashed horizontal lines indicate the RMSE of a model without features selection applied, therefore including all features with non-zero variance. Note that the x axis is cut at 500, whereas the total number of features with non-zero variance is higher for most models.

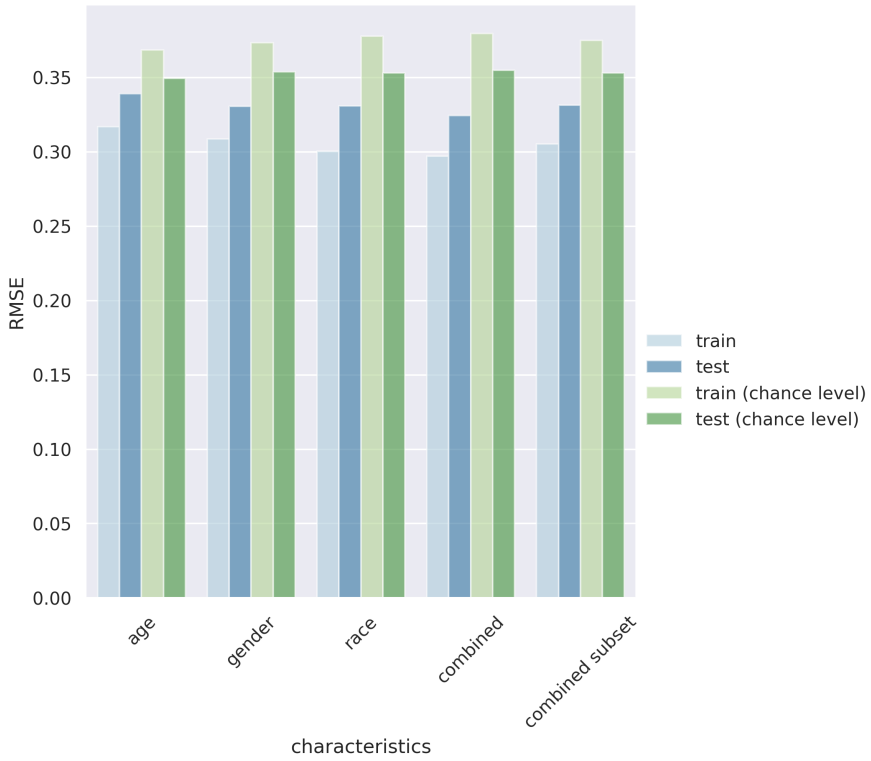

Figure 18: Testing errors for morph attack success prediction for different soft biometrics models. Along the x axis, different types of soft biometrics models have been used to extract embeddings for prediction. Either the embeddings of the age, gender, or race models have been used for prediction, or a combination of all three models, or an educated subset of most important features after feature selection. The y axis corresponds to the RMSE on predicting prodAvgMMPMR. Train and testing errors are illustrated in light and dark blue, respectively. A chance level of training and testing errors (model-dependent) is illustrated in light and dark green, respectively.

4 DISCUSSION

4.1 COSINE OR EUCLIDEAN DISTANCES ARE SUITABLE FOR IMAGE PRE-SELECTION

With track I (Section 2.1.1), different distance and similarity metrics have been evaluated for image pre-selection (Section 3.1). Cosine distance and Euclidean distance performed best, alongside a k-means constrained algorithm (Fig. 10). The presently used implementation of mutual information (Section 2.5.3) however did not perform that well, in terms that it did not level up prodAvgMMPMR as strongly as pre-selection based on Euclidean or Cosine distance did. However, resulting morphs still produced prodAvgMMPMR higher than random assignment. In the present implementation of mutual information (Section 2.5.3), a discretization step needed to be performed, and a number of 10 bins has been chosen for. Other numbers of bins have also been tested, but no qualitative difference was observed (not shown). Further, a continuous formulation of mutual information [Bel+18] has also been tested to avoid discretization, but the results were similar.

It must be noted however that the absolute results, i.e., the magnitudes of prodAvgMMPMRs are over-estimated. That is because the FRS used for pre-selection and the FRS used for verification were identical, namely the ArcFace model (Section 2.3.3). It is unlikely that an attacker would use by incidence the same system for potential pre-selection, as will be used in a real-world verification. Firstly, because the real-world verification FRS are rather closed-source commercial algorithms, and secondly by the sheer amount of different available FRSs. However, the analysis aimed at comparing pre-selection distances, so the relative relationship of the metrics' distributions is unlikely to differ if pre-selection and verification FRSs vary.

4.2 OPEN POTENTIAL FOR THE PRE-SELECTION ALGORITHM

Because Cosine distance performed slightly better than the competing pre-selection distances (i.e., Euclidean distance and Mutual Information score) and pre-selection methods (i.e., k-means constrained), for further analyses only the Cosine distance has been considered. Further, the k-means constrained algorithm would have run into troubles when more constrains were applied, such as there have been in track II. For instance, constrains can be set on how many samples (i.e., identities) were put into a leaf of the cluster tree. However, in its current implementation it would fail when more constrains were applied, such as age, gender and race constrains. To account for gender and race constrains, clustering could just be performed into subgroups of similar gender and race. However, to account for age, the data set

might have been split into many age groups to perform the clustering only within age (and gender, and race) subgroups. Possible matching pairs would be separated by group borders because an (arbitrary) cutoff on the age variable lied between them, as a result of the discretization of the variable.

Furthermore, more sophisticated algorithms could have been used for image pre-selection, instead of the naïve top down selection of pairs based on their similarity or distance, such as the Hungarian algorithm [Dro+21]. For instance Rottcher et al. used an algorithm based on minimum weight matching to optimize the distribution of the distance scores within the pairings of a data set [RSB20]. This algorithm would, similar than the k-means constrained algorithm above, be confronted with drawbacks such as the need for discretization of the age constrain. In a future version, the algorithm could be adjusted to penalize a high difference in a particular soft biometrics characteristics between the data subjects.

4.3 COMPARISON OF FACE RECOGNITION MODELS FOR PRE-SELECTION

Regarding the FRS for extracting embeddings for image pre-selection, different models have been evaluated. The results showed, that the recently published MagFace algorithms performed best, tightly followed by ArcFace. VGG-Face and in particular DeepFace showed relatively bad performance. However, all pre-selection methods improved the success of the morph attacks (Fig. 11, S3, 12, 13 & 14, Tab. 4 & 5). Further, a bias was seen, so that if the same FRS has been used for pre-selection as for verification, the FRS is more vulnerable to the resulting morphs (Fig. 11 & 13). However, using two COTS FRSs, this bias was not introduced and the pattern was still the same Fig. S3 & 14.

A – at first glance – counterintuitive observation can be made in Figure 11: Whereas *good* FRSs such as MagFace and ArcFace were quite vulnerable to morphing attacks, *bad* FRSs such as VGG-Face or DeepFace did not show any considerable vulnerability, as the prodAvgMMPMR when verified with those FRSs were mostly accumulating around 0. This trend is indicating, that by generally improving FRSs, so that after calibrating to a given FMR of 0.1% the FNMR becomes lower, these – in terms of recognition accuracy – *better* FRSs will become more vulnerable to morph attacks. The key figure is the decision threshold, which is located somewhere in between the distributions of the mated distances and the non-mated distances (Fig. 13). As long as a considerable proportion of the distribution of the morphed images is located below the threshold towards the mated comparisons, the FRS will be quite vulnerable. Adjusting the decision threshold towards the distribution of the mated comparisons would diminish this vulnerability. Adjusting that decision threshold would best be possible in MagFace as verification model, as the mated and morphed distributions showed a small overlap (Fig. 13). With a model as good as ArcFace as well as the two COTS FRSs, the distributions were however already showing a considerable overlap, impeding an simple solution via adjustment. Furthermore, by adjusting the decision

threshold in the direction of the mated comparisons, FMR would decrease, which in general makes the system more secure – also against zero-effort impostor attacks. This in turn would inevitably increase the FNMR, rendering the system less convenient for practical purpose. In addition, the presently used morphs were produced in an automated fashion. A real world attacker would be able to invest time and resources into creating one single, high quality morph, by manual intervention and various image post-processing steps. Comparison scores by such manually created morphs would be even more challenging to distinguish from mated comparisons, even when using MagFace for verification.

From the distribution of the prodAvgMMPMRs in Figure 11 – the high vulnerability of MagFace and ArcFace and the low vulnerability of VGG-Face and DeepFace – some inference on the results of the MVRs can be drawn (Fig. 12). In particular, the high values of the four leftmost columns in each MVR matrix are likely to derive from the more vulnerable MagFace and ArcFace FRSs, and the two COTS FRSs. Analogously, the quite low values in the two rightmost columns are likely to be driven be the less vulnerable FRSs DeepFace and VGG-Face.

When correcting the mated morph rates for the FNMR of a verification FRS as has been done using the RMMR metric (eq. 7, Tab. 4), the general pattern persisted that a verification FRS was most vulnerable to morphs from image pairs pre-selected with the embeddings of the identical FRS. However, by ranking the RMMR row-wise and average across pre-selection methods and morphing algorithms (Tab. 5), the pattern manifests that MagFace is best suited for pre-selection among the tested FRSs. ArcFace follows MagFace, then VGG-Face, and lastly DeepFace. Poorest performance has constantly been seen by randomly pre-selected morphs.

4.4 EVALUATION OF MORPHING ALGORITHMS FOR PRE-SELECTION

A clear performance gap between morphing algorithms runs like a common thread through all analyses. Morphed images created by UBO morpher, closely followed by those morphed by NTNU morpher, performed best in fooling both FRSs (Fig. 10 & 11 and Tab. 4) and also D-MAD algorithms (Fig. 15 & 16). Morphs created by Alyssaq morpher and MIPGAN however performed worse in the present analyses. As especially Alyssaq morpher is extremely effective in terms of computational time, it can still be used as invaluable tool to generate a large amount of morphed images, as has been done e.g., for track III.

What can be seen from Figure 15 and Figure 16 is that the morphing algorithm deployed has a higher impact on the success of the D-MAD algorithm, than the pre-selection. Similar accounts for the success in terms of fooling the verification FRSs, as can be seen in Figure 11, and by comparing the MVRs between morphers (Fig. 12, S4, S5, & S6). Alyssaq and MIPGAN morphers performed rather low in fooling the D-MAD algorithm, even with pre-selection applied. The reason for Alyssaq morpher might for instance be the shape of

the resulting morph (Fig. 7). Alyssaq morpher returned morphs which were cropped at the facial borders in a non-rectangular fashion (Fig. 7), and not projected back to one of the original images' backgrounds. This has probably helped the D-MAD algorithm in its decision both during training and testing. Real world attackers would not use such a morph i.e. in a passport fraud scenario. Furthermore, MIPGAN morpher produced very blurry images (Fig. 7). In the original implementation of MIPGAN [Zha+21], the morphs were of higher quality, but also the original images used for morphing were of better image quality than the database used in the study at hand. The morphing in the latent space might therefore in the present case have dropped many facial characteristics, that might have been helpful to facilitate a morph attack.

4.5 MORPH ATTACK DETECTION PERFORMANCE

Instead of adjusting decision thresholds to counter morphing attacks as proposed in Section 4.3, MAD algorithms might be interposed in a face verification process. The presently used D-MAD algorithm was introduced by [Sch+20] and learned to differentiate between the distribution of the differences of two bona fide images and the distribution of differences of morphs and bona fide images (Fig. 9, Section 2.11), all in the embedding space.

Testing on random morphing produced lowest BPCER10 values, indicating highest accuracy (Fig. 15). Testing on the other morphs increased BPCER10 values. Therefore, highest vulnerability of the D-MAD classifier has been seen for morphs pre-selected by MagFace, then ArcFace, VGG-Face, and lastly DeepFace. This was irrespective of whether the classifier was trained and tested with ArcFace embeddings or with MagFace embeddings.

In fact, the D-MAD algorithm trained with MagFace embeddings revealed considerable lower BPCER10 values, irrespective of the type of pre-selection. Therefore, using MagFace instead of ArcFace might be a significant improvement to the D-MAD classifier proposed by Scherhag et al. [Sch+20]. Please note that only the embeddings of the MagFace algorithm have been used and not an additional quality metrics returned by the model. However, the quality of an image was still incorporated in the embeddings by the way the loss function was constructed. In MagFace's loss function, high quality samples of an individual are drawn towards the center of the multidimensional distribution, whereas the low quality samples are pushed towards its borders [Men+21]. In other words, during training of MagFace, the magnitude of the face embeddings were made proportional to the Cosine distance to the respective class (i.e., individuals) centers [Fu+21]. Therefore, by having different image qualities for the bona fide images and the morphed images results in an easier separation of the both groups by the classifier, as their positions in 512-dimensional embedding space are farther apart than the positions of two high quality bona fide images.

One aims of large-scale image pre-selection based on embeddings was to evaluate a method for providing a sufficiently large data set of morphed face images to train MAD algorithms. Interestingly, a recent study showed,

that for training MAD algorithms, image pre-selection might be done in the opposite fashion than in the present study [Dam+19]. They demonstrated, that that training morphing pairs with low similarity can increase the MAD algorithm's performance [Dam+19].

In the study at hand, separate D-MAD algorithm were trained on either ArcFace or MagFace embeddings. However, a fusion of both might result in constructive effects. In particular, it is not yet shown if the combination of both D-MAD algorithms would perform better than using MagFace embeddings alone. Moreover, if subsets of the embeddings of both (and other FRSs) embedding vectors might be used. This could similarly improve computation time for pre-selection by at the same time reaching high performance. On the other hand, extracting embeddings of several FRSs adds considerable computation time in the first place. Moreover, if feature selection for dimensionality reduction on face embeddings adds as high value than in other use cases is still open. That is because the face recognition models learn the, e.g. 512-dimensional, representations, and removing any dimension might lead to a drop in accuracy, assuming that the variance in each dimension is equal. In that case, the whole might be more then the sum of its parts.

4.6 SOFT BIOMETRICS AND MORPH ATTACK SUCCESS

Previous studies have already used image pre-selection based on soft biometrics such as age, gender, or race (e.g., [Raj+20]). In the setting of a real world morphing attack, pre-selection based on these soft biometrics make sense as there expression is often linked to the issued identification document, and deviations might be suspicious to the authority issuing the document. Further, the resulting morph might appear less authentic when morphing was performed across gender or race, or with high age difference.

The results of the multilevel model (Tab. 6) demonstrate differences in the morph attack potential related to the expression of the soft biometrics characteristics, even if they closely match between individuals. For instance, increasing age diminished the morph attack success, as well as the affiliation to ethnicity. Contrary, gender did not play a role for morph success. However, the data set is very unbalanced according to soft biometrics characteristics. Results must be confirmed by in-depth analysis of more balanced data sets such as for example FairFace [KJ21]. FairFace however had the disadvantage of comprising faces-in-the-wild, which introduces an abundance of artifacts in the resulting morphs. Similar data sets with more controlled and ICAO compliant [Int15] images might however be used.

Using embeddings of soft biometrics model to predict morph attack success was less promising (Section 3.3). The embeddings only slightly decreased the error on the prediction of the MMPMRs compared to chance level (Fig. 18). It was however shown, that a small subset of the soft biometrics embeddings was sufficient to predict morph success to a similar level, but with generally still a high error and a slightly higher error than without fea-

ture selection (Fig. 17). However, reducing the number of features can still be useful to speed up calculation time for pre-selection within large data sets.

The main issue in this analysis might however be the data foundation in the first place. The activations of a middle layer of each soft biometrics model were extracted as embeddings. Activations were of very heterogeneous distributions, and not easily to normalize to a convenient distribution, which typical statistical models would have been possible to be applied to. Therefore, a tree-based algorithm (Section 2.12.2) has been used, which were more naïve to the heterogeneous data distribution by using multiple, binary splits of the multidimensional data space.

Moreover, the three different types of soft biometrics embeddings all performed similarly, with only small deviations (Fig. 18). This might in part be related to the fact, that all derived from the same original model, the VGG-Face model (Fig. 4). Although the three soft biometrics models all returned the same number of embeddings, the number of embeddings with non-zero variance across observations was very different, with the age model returning around 500, the gender around 1000, and the race model around 2000 potential features. Combining features from all models did not improve morph attack success prediction. This kind of analysis should however be reinvestigated with embeddings of more elaborate soft biometrics models than the presently used.

The weak performance of the approach to use soft biometrics embeddings to enhance image pre-selection for morphing however lines up to related findings from earlier work [RSB20]. There, similarity in characteristics such as age, skin tone, or hair shape has been used as potential features for pre-selection. However, that method for image pre-selection was outperformed by the approach of pre-selection using face embeddings [RSB20].

4.7 CONCLUSION

The study at hand conducted a detailed analysis of face embeddings in the context of image pre-selection for morphing. Face embeddings were highly suitable for image pre-selection, especially when MagFace or ArcFace embeddings have been used. Furthermore, MagFace embeddings turned out to be particular useful to increase performance of D-MAD, and offer new potential for further research in this field.

Part II

APPENDIX

A

SUPPLEMENTARY METHODS

A.1 DET CURVES OF VERIFICATION FRS

During calculation of the verification thresholds for the FRGCv2 (Fig. S1) and the UNCW (Fig. S2) data set (Section 2.8), each FRS was tested with several genuine and impostor comparisons.

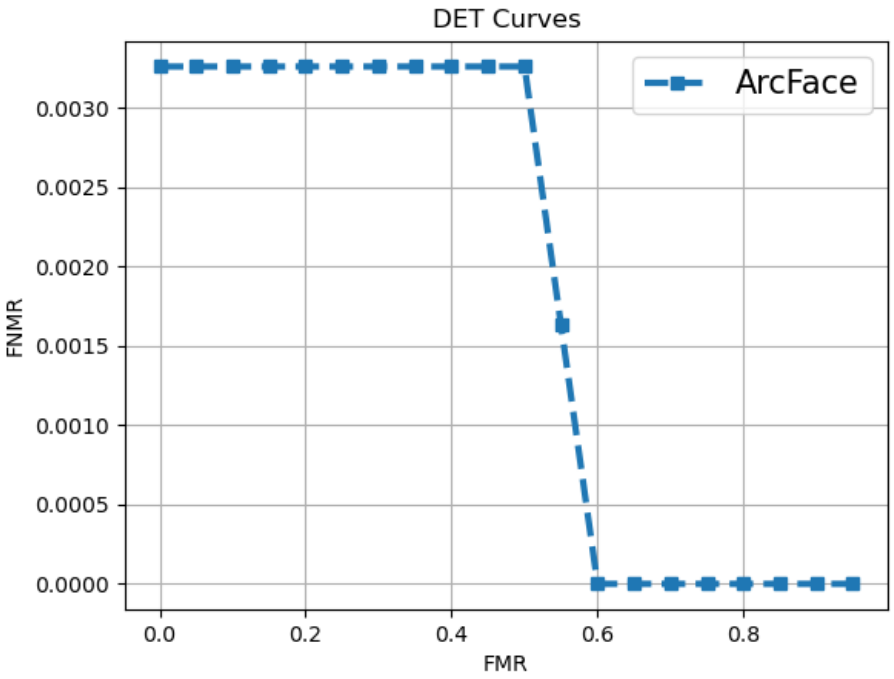

Figure S1: DET curves for the used open-source FRS on the FRGCv2 data set. The FNMR (y axis) is plotted against the FMR (x axis), both of which were determined by varying the verification threshold τ of the FRS.

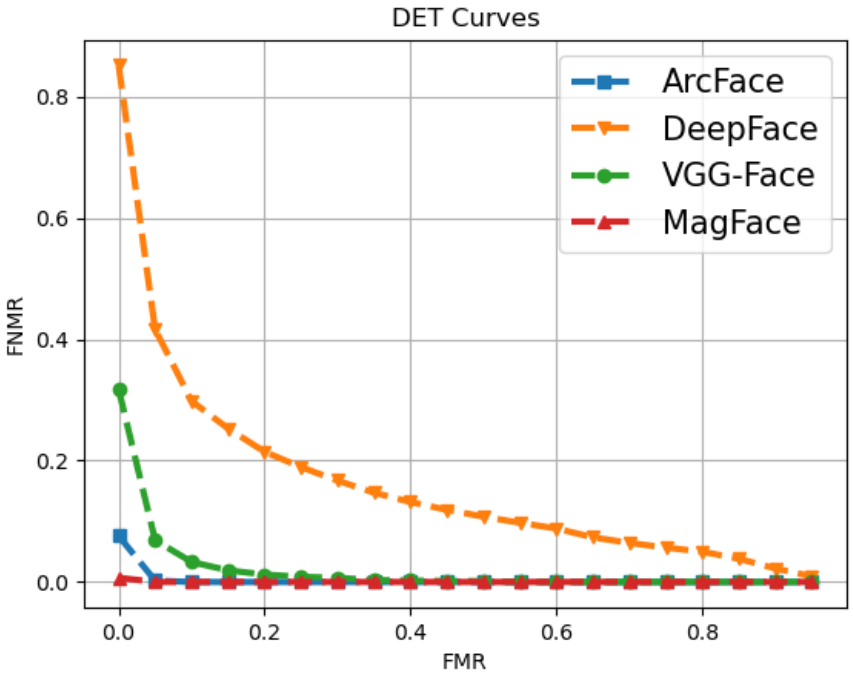

Figure S2: DET curves for the four used open-source FRSs on the UNCW data set. The FNMR (y axis) is plotted against the FMR (x axis), both of which were determined by varying the verification threshold τ of the FRS. The FRSs are illustrated in different colors. The best overall performance – in terms of both low FMR and low FNMR – showed MagFace, followed by ArcFace, VGG-Face, and lastly DeepFace.

B

SUPPLEMENTARY RESULTS

B.1 MATED MORPH PRESENTATION MATCH RATES ON COTS FRSS

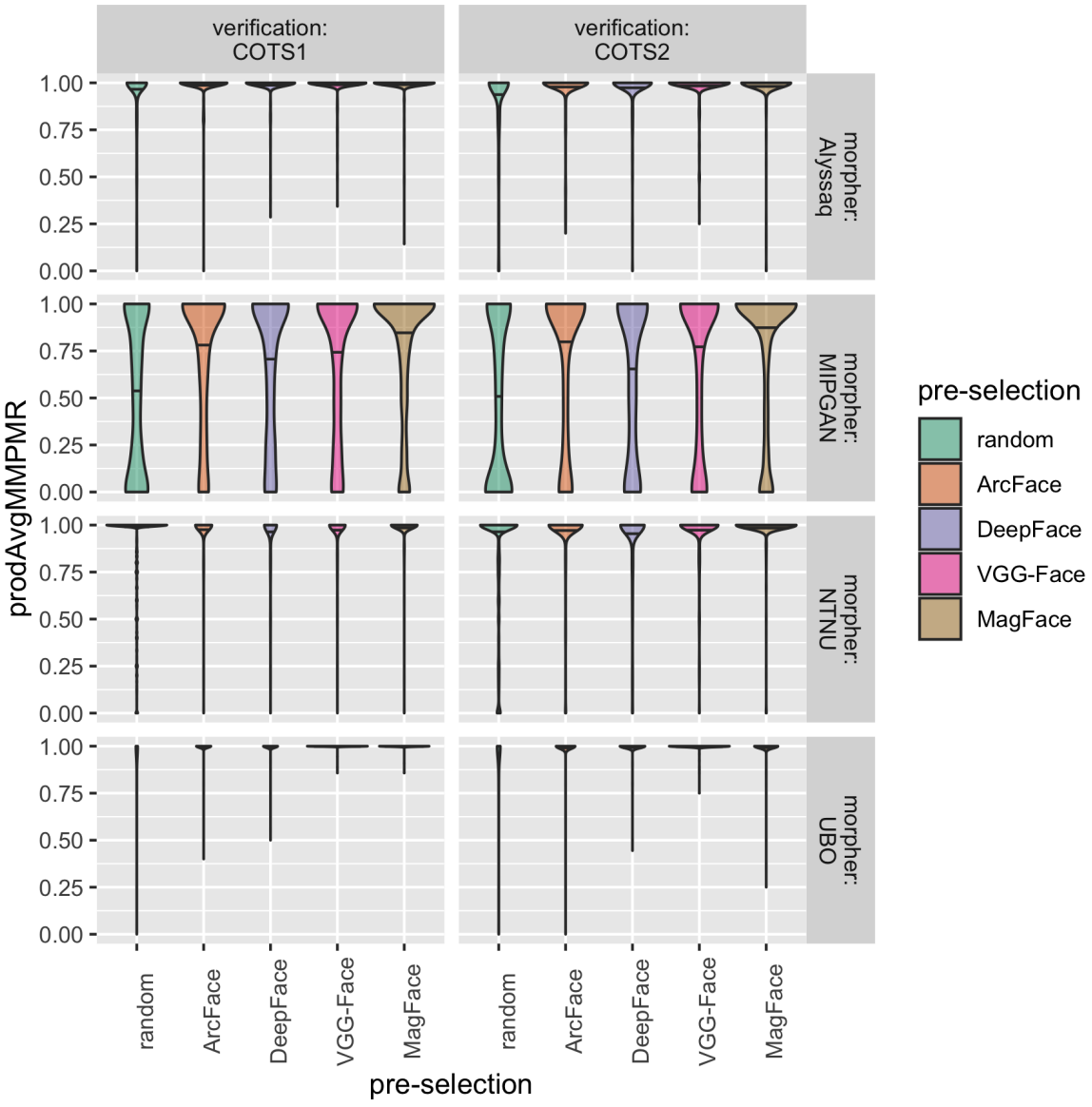

Figure S3: Mated morphs comparison success rates for different image pre-selection embeddings. prodAvgMMPMRs (y axes) are plotted for different pre-selection methods (x axis & color-coded). Density is plotted in horizontal direction. Median values are illustrated by horizontal black bars. The same pairs have been morphed by different morphing methods (rows). Random assignment of the morphing pairs are displayed in the left-most column. All morphs have been evaluated by different COTS FRSs (columns). Note that prodAvgMMPMR was calculated as fraction within range $[0; 1]$.

B.2 MORPH VULNERABILITY RATES (MVR)

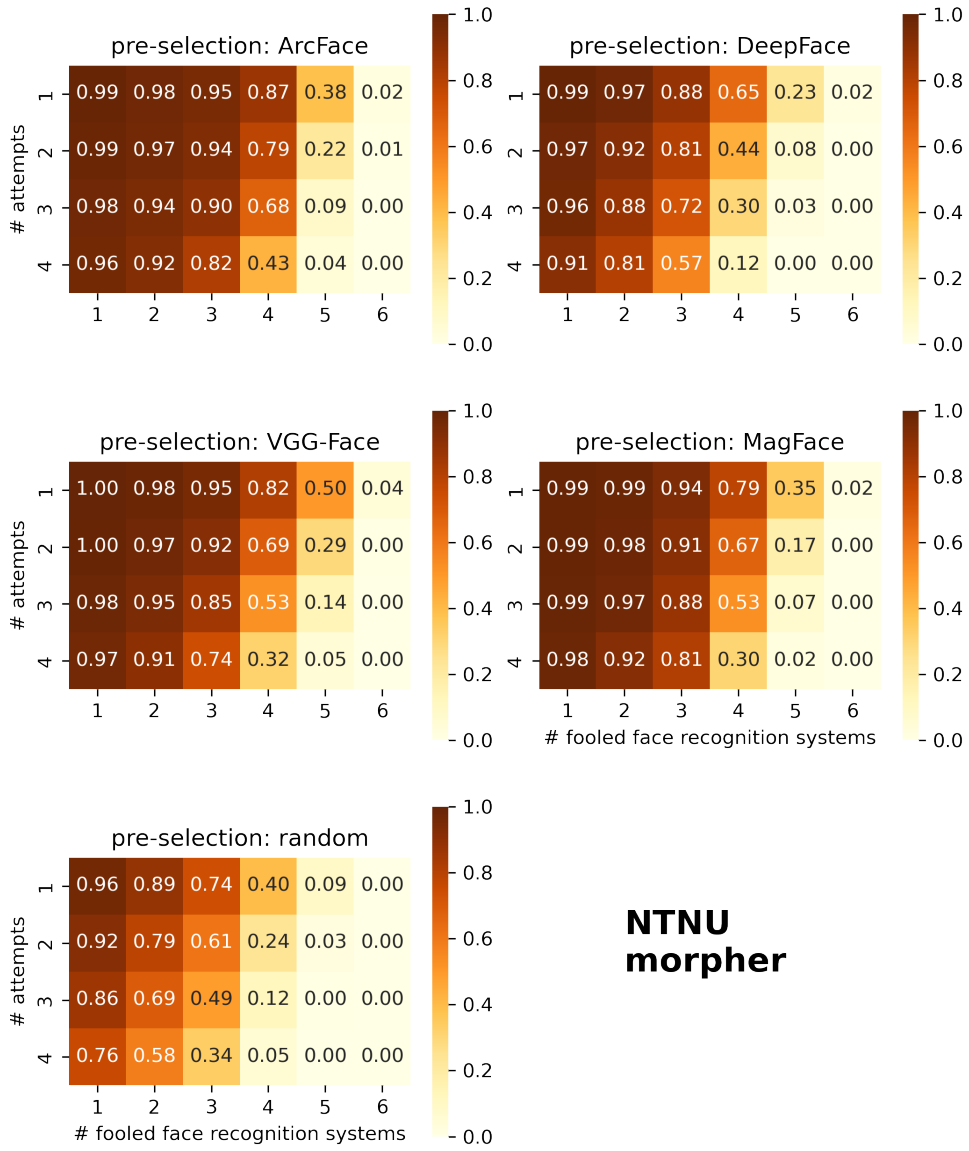

Figure S4: MVR of morphs generated by NTNU morpher. See Figure 12 for details.

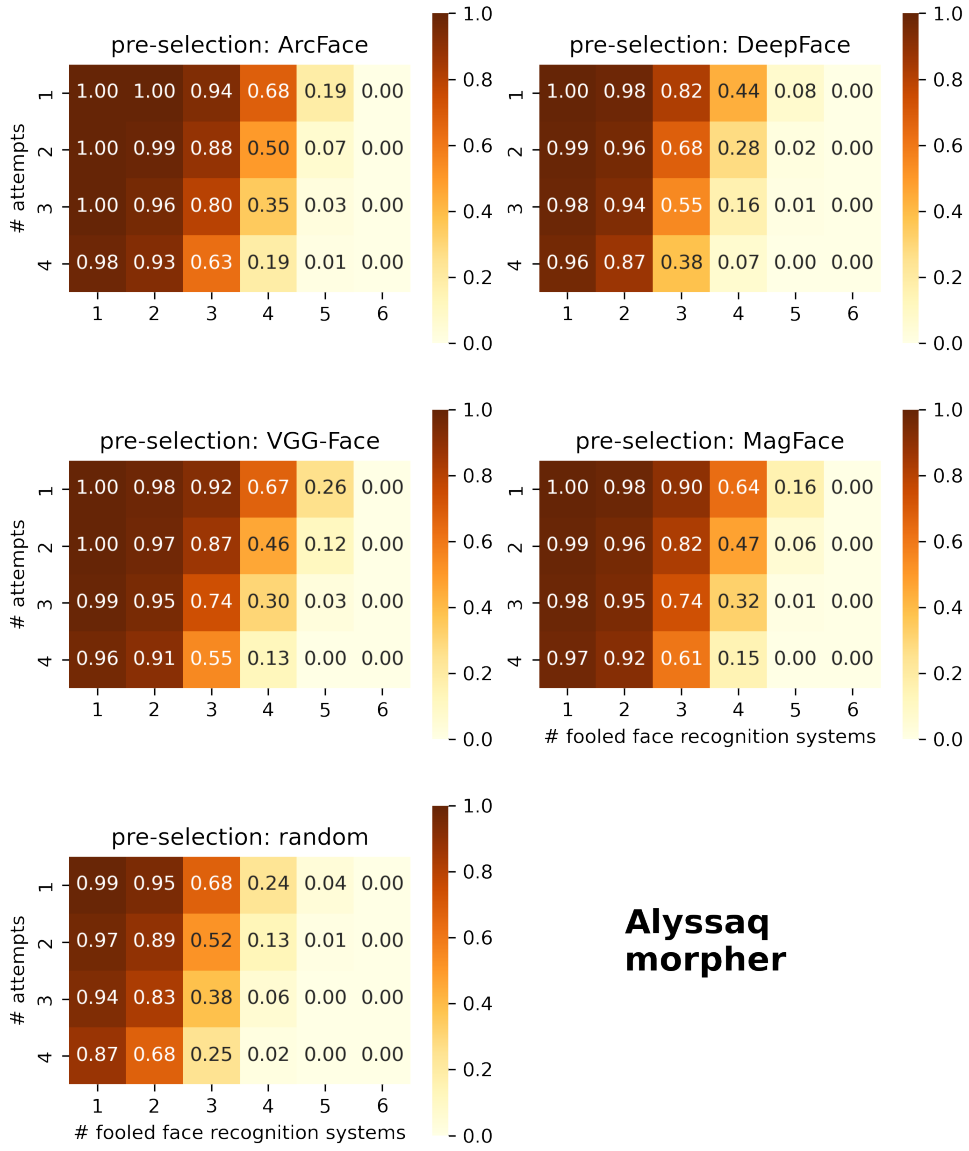

Figure S5: MVR of morphs generated by Alyssaq morpher. See Figure 12 for details.

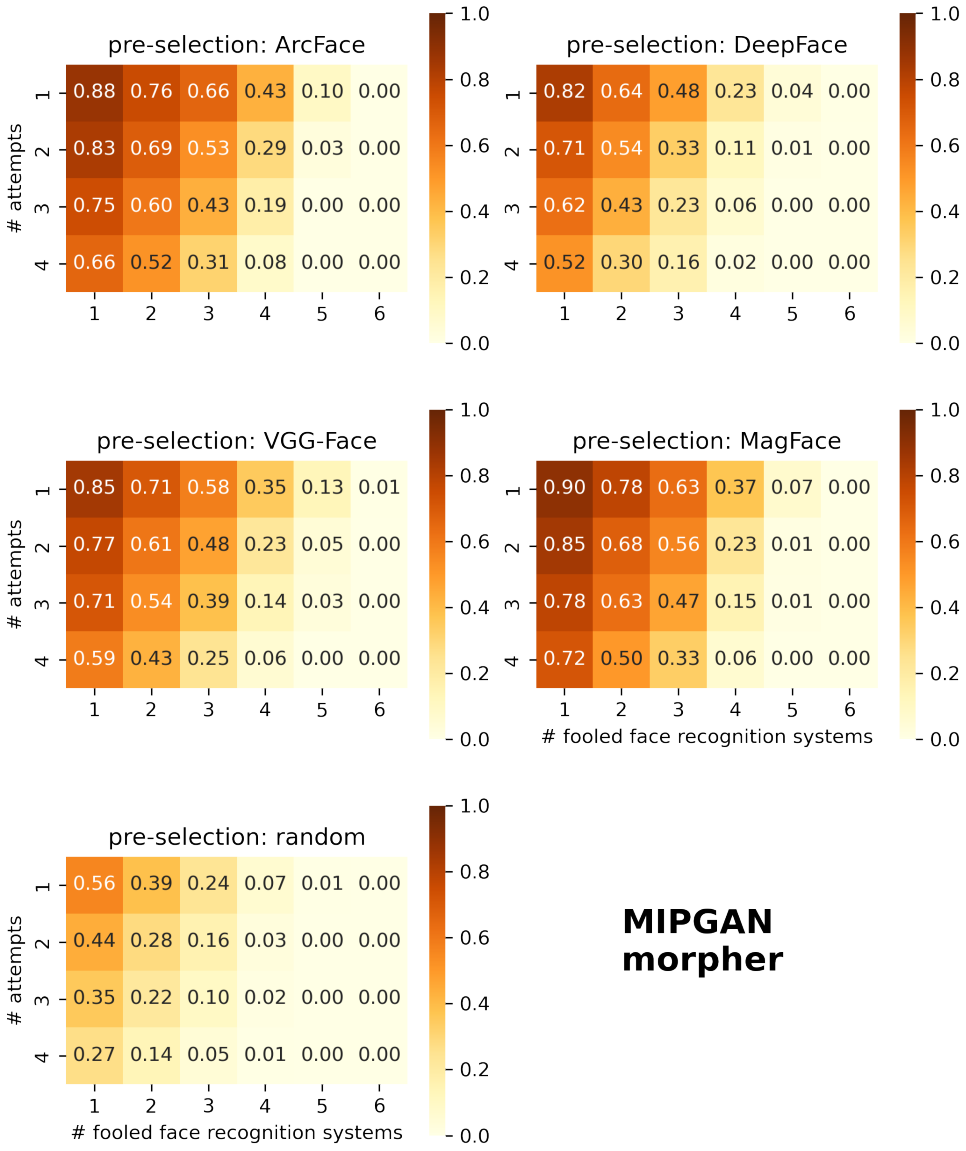

Figure S6: MVR of morphs generated by MIPGAN morpher. See Figure 12 for details.

B.3 DISTRIBUTION OF MATED MORPH COMPARISON SCORES (OPEN SOURCE FRSS)

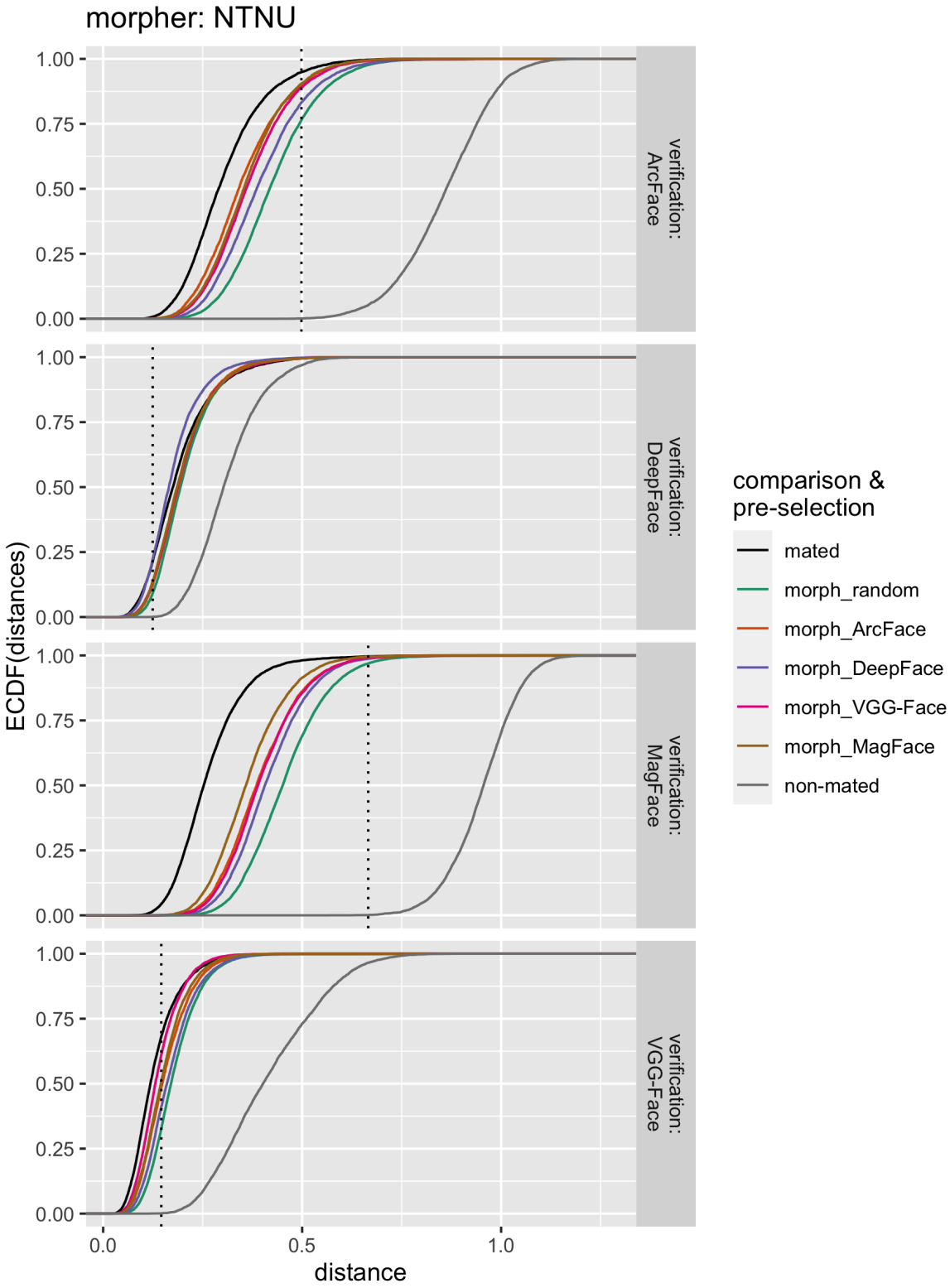

Figure S7: ECDFs for distance scores of the open source FRSs. Morphs were created by NTNU morpher. See Figure 13 for details.

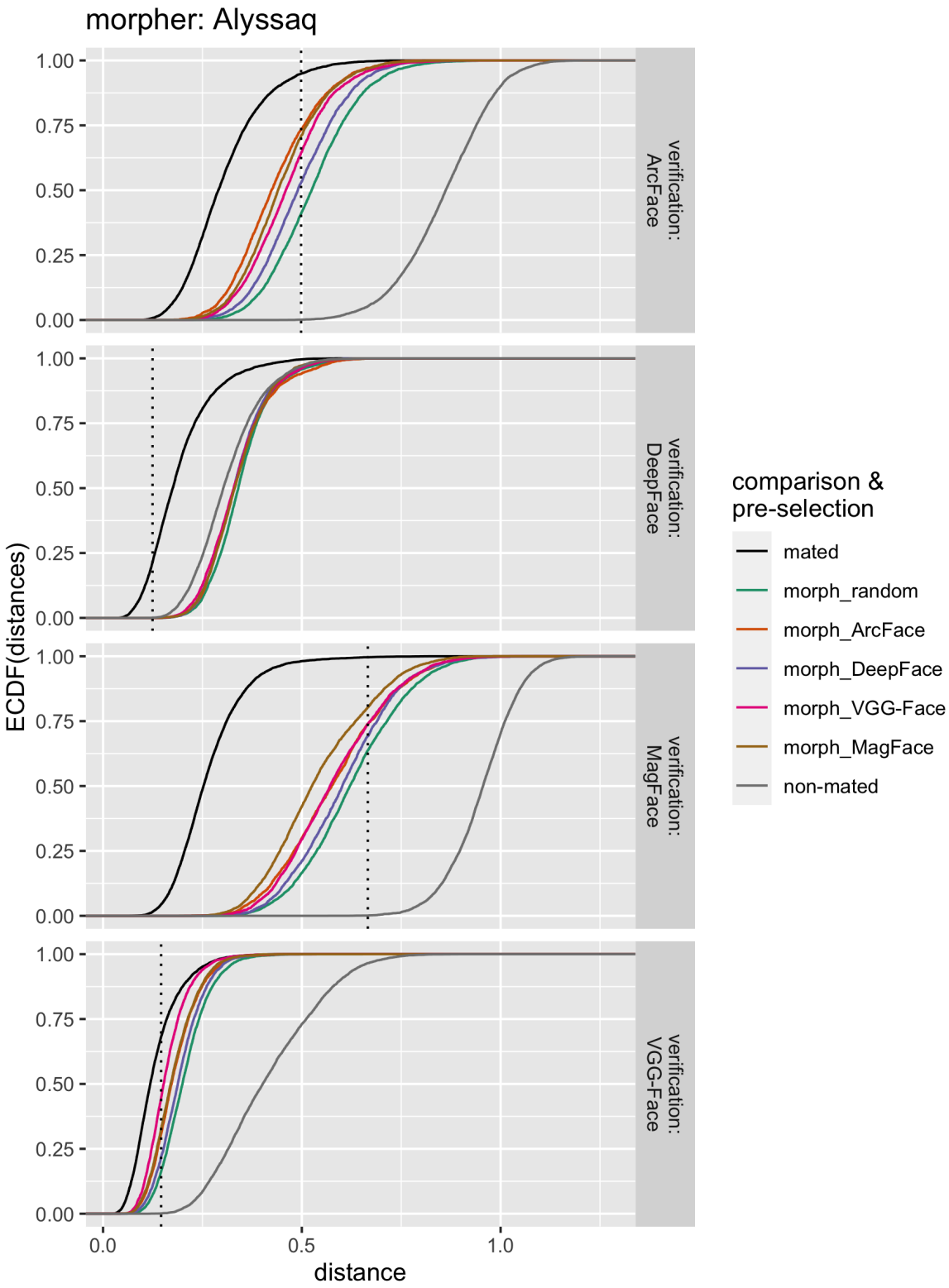

Figure S8: ECDFs for distance scores of the open source FRSs. Morphs were created by Alyssaq morpher. See Figure 13 for details.

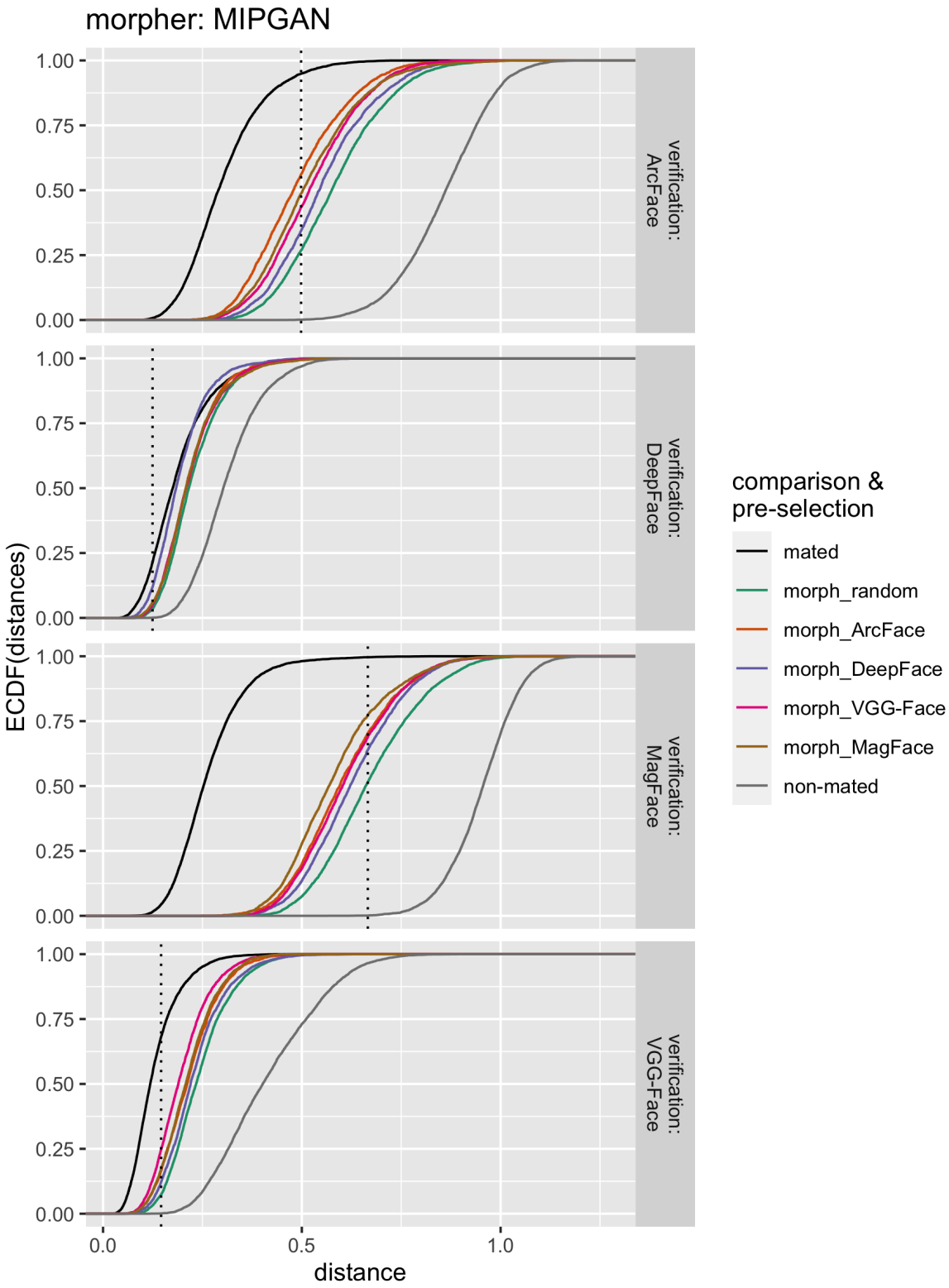

Figure S9: ECDFs for distance scores of the open source FRSs. Morphs were created by MIPGAN morpher. See Figure 13 for details.

B.4 DISTRIBUTION OF MATED MORPH COMPARISON SCORES (COTS FRSS)

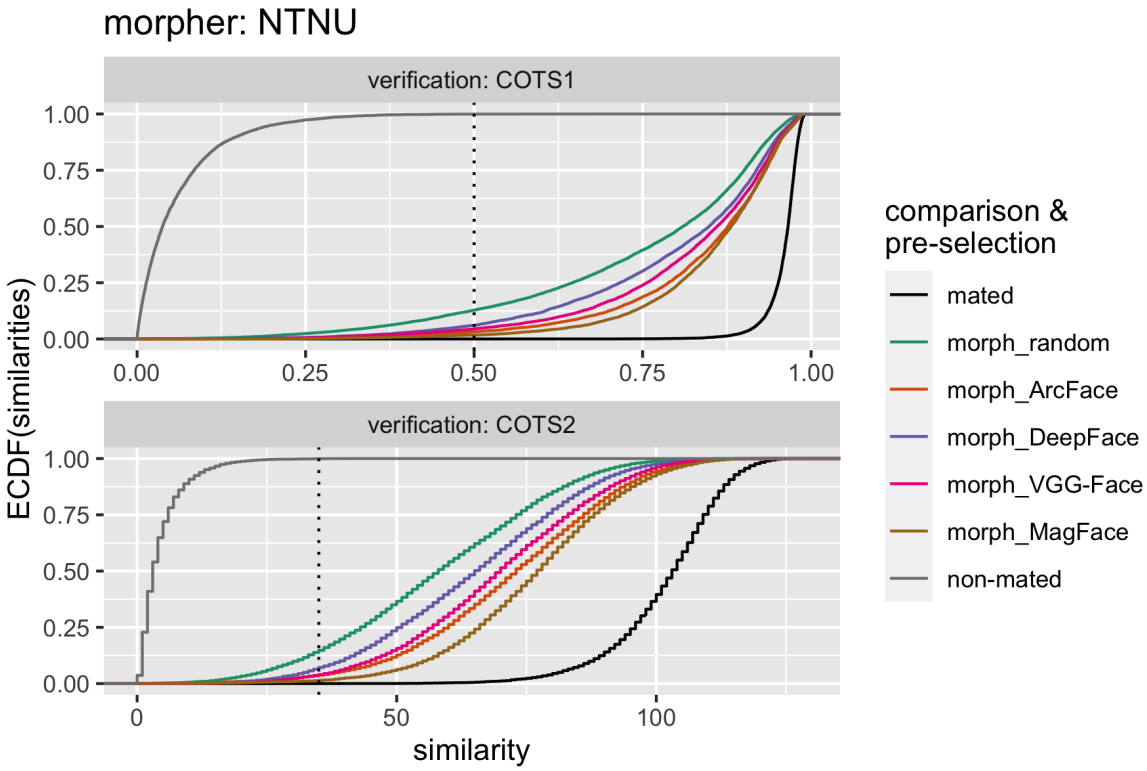

Figure S10: ECDFs for similarity scores of the COTS FRSs. Morphs were created by NTNU morpher. See Figure 14 for details.

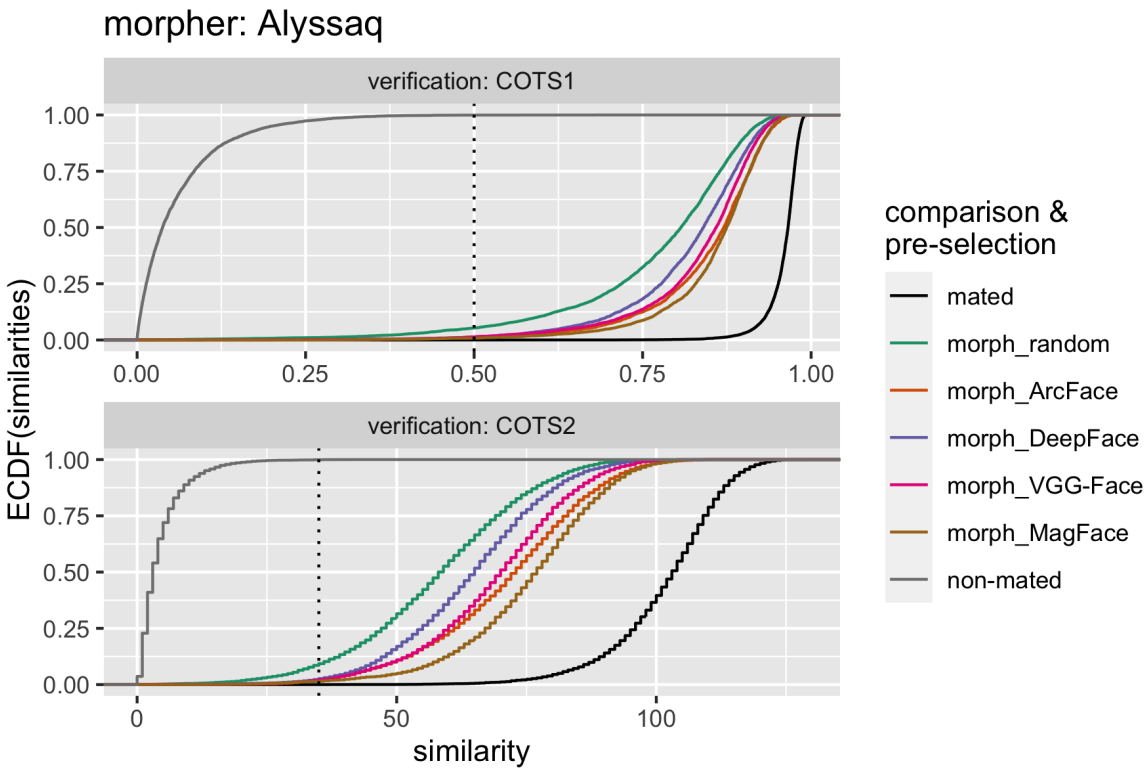

Figure S11: ECDFs for similarity scores of the COTS FRSs. Morphs were created by Alyssaq morpher. See Figure 14 for details.

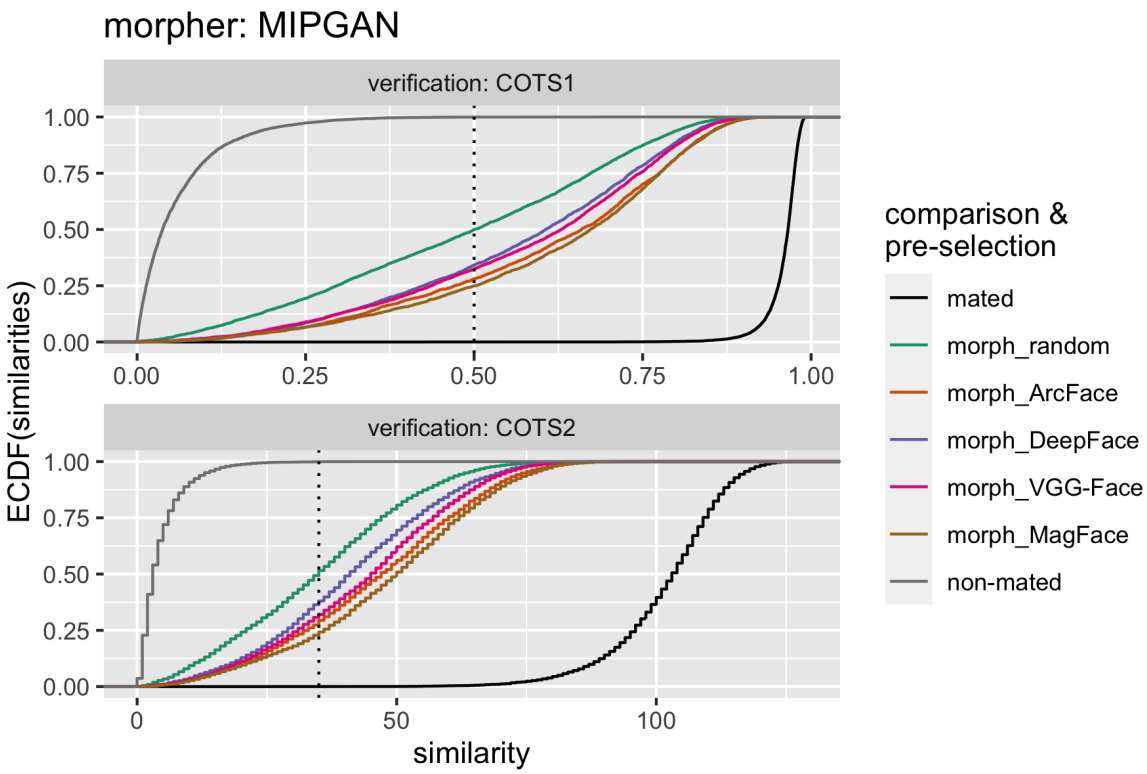

Figure S12: ECDFs for similarity scores of the COTS FRSs. Morphs were created by MIPGAN morpher. See Figure 14 for details.

BIBLIOGRAPHY

[AMAZ17] Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. "Understanding of a convolutional neural network." In: *2017 International Conference on Engineering and Technology (ICET)*. 2017 International Conference on Engineering and Technology (ICET). Antalya: IEEE, Aug. 2017, pp. 1–6.

[ALS] Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. "OpenFace: A general-purpose face recognition library with mobile applications." In: *CMU School of Computer Science* (), p. 20.

[Bat+15] Douglas Bates, Martin Mächler, Benjamin M. Bolker, and Steven C. Walker. "Fitting linear mixed-effects models using lme4." In: *Journal of Statistical Software* 67.1 (2015).

[Bel+18] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. "Mine: mutual information neural estimation." In: *arXiv preprint arXiv:1801.04062* (2018).

[BBD00] Paul S Bradley, Kristin P Bennett, and Ayhan Demiriz. "Constrained k-means clustering." In: *Microsoft Research, Redmond* 20.0 (2000), p. 0.

[Bra00] G. Bradski. *The OpenCV library*. 2000.

[Dam+19] Naser Damer, Alexandra Mosegui Saladie, Steffen Zienert, Yaza Wainakh, Philipp Terhörst, Florian Kirchbuchner, and Arjan Kuijper. "To detect or not to detect: The right faces to morph." In: *2019 international conference on biometrics (ICB)*. 2019, pp. 1–8.

[Den+19] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. "ArcFace: Additive Angular Margin Loss for Deep Face Recognition." In: *arXiv:1801.07698 [cs]* (Feb. 9, 2019). arXiv: 1801.07698.

[DP05] Chris Ding and Hanchuan Peng. "Minimum Redundancy Feature Selection from Microarray Gene Expression Data." In: *Journal of bioinformatics and computational biology* (2005), p. 21.

[Dro+21] Pawel Drozdowski, Fabian Stockhardt, Christian Rathgeb, Daile Osorio-Roig, and Christoph Busch. "Feature fusion methods for indexing and retrieval of biometric data: Application to face recognition with privacy protection." In: *IEEE access : practical innovations, open solutions* 9 (2021). Publisher: IEEE, pp. 139361–139378.

[FFM14] Matteo Ferrara, Annalisa Franco, and Davide Maltoni. "The magic passport." In: *IEEE international joint conference on biometrics*. 2014, pp. 1–7.

[FFM16] Matteo Ferrara, Annalisa Franco, and Davide Maltoni. "On the effects of image alterations on face recognition accuracy." In: *Face recognition across the imaging spectrum*. 2016, pp. 195–222.

[FFM18] Matteo Ferrara, Annalisa Franco, and Davide Maltoni. "Face demorphing." In: *IEEE Transactions on Information Forensics and Security* 13.4 (2018), pp. 1008–1017.

[FFM19] Matteo Ferrara, Annalisa Franco, and Davide Maltoni. "Decoupling texture blending and shape warping in face morphing." In: *2019 international conference of the biometrics special interest group (BIOSIG)*. 2019, pp. 1–5.

[FFM21] Matteo Ferrara, Annalisa Franco, and Davide Maltoni. "Face morphing detection in the presence of printing/scanning and heterogeneous image sources." In: *IET Biometrics* 10.3 (May 2021), pp. 290–303.

[Fer+22] Matteo Ferrara, Annalisa Franco, Davide Maltoni, and Christoph Busch. "Morph Vulnerability Rate." In: *(in process)* (2022).

[Fri01] Jerome H Friedman. "Greedy function approximation: a gradient boosting machine." In: *Annals of statistics* (2001). Publisher: JSTOR, pp. 1189–1232.

[Fro15] Frontex. *Best practice technical guidelines for automated border control (ABC) systems*. 2015.

[Fu+21] Biying Fu, Noémie Spiller, Cong Chen, and Naser Damer. "The effect of face morphing on face image quality." In: *2021 international conference of the biometrics special interest group (BIOSIG)*. 2021, pp. 1–5.

[ISO17] ISO/IEC. *International Standard ISO/IEC 30107-3 (First edition) Information technology - Biometric presentation attack detection - Part 3: Testing and reporting*. 2017.

[ISO21] ISO/IEC. *International Standard ISO/IEC 19795-1 (Second edition) Information technology - Biometric performance testing and reporting - Part 1: Principles and frameworks.* 2021.

[Int15] International Civil Aviation Organization. *ICAO doc 9303, machine readable travel documents – Part 9: deployment of biometric identification and electronic storage of data in MRTDs*. 2015.

[KJ21] Kimmo Karkkainen and Jungseock Joo. "FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation." In: *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2021 IEEE Winter Conference on Applications of Computer Vision (WACV). Waikoloa, HI, USA: IEEE, Jan. 2021, pp. 1547–1557.

[KLA19] Tero Karras, Samuli Laine, and Timo Aila. "A Style-Based Generator Architecture for Generative Adversarial Networks." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), p. 10.

[Kar+20] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. "Analyzing and Improving the Image Quality of StyleGAN." In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, June 2020, pp. 8107–8116.

[KS14] Vahid Kazemi and Josephine Sullivan. "One millisecond face alignment with an ensemble of regression trees." In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Columbus, OH: IEEE, June 2014, pp. 1867–1874.

[Ke+17] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. "Lightgbm: A highly efficient gradient boosting decision tree." In: *Advances in neural information processing systems* 30 (2017), pp. 3146–3154.

[Kor+20] Yassin Kortli, Maher Jridi, Ayman Al Falou, and Mohamed Atri. "Face Recognition Systems: A Survey." In: *Sensors* 20.2 (Jan. 7, 2020), p. 342.

[LB90] Mary J Lindstrom and Douglas M. Bates. "Nonlinear Mixed Effects Models for Repeated Measures Data." In: *Biometrics* 46.3 (1990), pp. 673–687.

[Mal13] MD Malkauthekar. "Analysis of Euclidean distance and Manhattan distance measure in Face recognition." In: *Third international conference on computational intelligence and information technology (CIIT 2013)*. tex.organization: IET. 2013, pp. 503–507.

[Mar] Manuel Aguado Martinez. *PyEER*. URL: https://github.com/manuelaguadomtz/pyeer.

[Men+21] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. "MagFace: A Universal Representation for Face Recognition and Quality Assessment." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), p. 10.

[NB10] Hieu V Nguyen and Li Bai. "Cosine similarity metric learning for face verification." In: *Asian conference on computer vision*. 2010, pp. 709–720.

[NCW] University of North Carolina Wilmington. *MORPH Non-Commercial Release Whitepaper*. URL: http://people.uncw.edu/vetterr/MORPH-NonCommercial-Stats.pdf.

[PVZ15] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. "Deep Face Recognition." In: *Procedings of the British Machine Vision Conference 2015*. British Machine Vision Conference 2015. Swansea: British Machine Vision Association, 2015, pp. 41.1–41.12.

[Phi+05] P Jonathon Phillips, Patrick J Flynn, Todd Scruggs, Kevin W Bowyer, Jin Chang, Kevin Hoffman, Joe Marques, Jaesik Min, and William Worek. "Overview of the Face Recognition Grand Challenge." In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR05). Vol. 1. San Diego, CA, USA: IEEE, 2005, pp. 947–954.

[Que] Alyssa Quek. *face morpher*. URL: https://github.com/alyssaq/face_morpher.

[Rag+17] Ramachandra Raghavendra, Kiran Raja, Sushma Venkatesh, and Christoph Busch. "Face morphing versus face averaging: Vulnerability and detection." In: *2017 IEEE International Joint Conference on Biometrics (IJCB)*. 2017 IEEE International Joint Conference on Biometrics (IJCB). Denver, CO: IEEE, Oct. 2017, pp. 555–563.

[Raj+20] Kiran Raja et al. "Morphing Attack Detection – Database, Evaluation Platform and Benchmarking." In: *arXiv:2006.06458 [cs]* (Sept. 28, 2020). arXiv: 2006.06458.

[Rio+16] Jose Sanchez del Rio, Daniela Moctezuma, Cristina Conde, Isaac Martin de Diego, and Enrique Cabello. "Automated border control e-gates and facial recognition systems." In: *Computers & Security* 62 (Sept. 2016), pp. 49–72.

[RSB20] Alexander Roettcher, Ulrich Scherhag, and Christoph Busch. "Finding the Suitable Doppelgänger for a Face Morphing Attack." In: *2020 IEEE International Joint Conference on Biometrics (IJCB)*. 2020 IEEE International Joint Conference on Biometrics (IJCB). Houston, TX, USA: IEEE, Sept. 28, 2020, pp. 1–7.

[Sch90] Robert E Schapire. "The strength of weak learnability." In: *Machine learning* 5.2 (1990). Publisher: Springer, pp. 197–227.

[Sch+16] Ulrich Scherhag, Christian Rathgeb, Johannes Merkle, Ralph Breithaupt, and Christoph Busch. "Face Recognition Systems Under Morphing Attacks: A Survey." In: *IEEE Access* 7 (2016), pp. 23012–23026.

[Sch+20] Ulrich Scherhag, Christian Rathgeb, Johannes Merkle, and Christoph Busch. "Deep Face Representations for Differential Morphing Attack Detection." In: *arXiv:2001.01202 [cs]* (Apr. 3, 2020). arXiv: 2001.01202.

[Sch+17] Ulrich Scherhag et al. "Biometric Systems under Morphing Attacks: Assessment of Morphing Techniques and Vulnerability Reporting." In: *2017 International Conference of the Biometrics Special Interest Group (BIOSIG)*. 2017 International Conference of the Biometrics Special Interest Group (BIOSIG). Darmstadt, Germany: IEEE, Sept. 2017, pp. 1–7.

[SKP15] Florian Schroff, Dmitry Kalenichenko, and James Philbin. "FaceNet: A Unified Embedding for Face Recognition and Clustering." In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2015), pp. 815–823. arXiv: 1503.03832.

[SO20] Sefik Ilkin Serengil and Alper Ozpinar. "LightFace: A hybrid deep face recognition framework." In: *2020 innovations in intelligent systems and applications conference (ASYU)*. 2020, pp. 23–27.

[SJ19] Yichun Shi and Anil Jain. "Probabilistic Face Embeddings." In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE, Oct. 2019, pp. 6901–6910.

[Tai+14] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. "DeepFace: Closing the Gap to Human-Level Performance in Face Verification." In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Columbus, OH, USA: IEEE, June 2014, pp. 1701–1708.

[To13] R Core Team and others. *R: A language and environment for statistical computing*. 2013.

[UNC] UNCW. *MORPH Non-Commercial Release Whitepaper.* URL: http://people.uncw.edu/vetterr/MORPH-NonCommercial-Stats.pdf (visited on 12/03/2021).

[Ven+21] Sushma Venkatesh, Raghavendra Ramachandra, Kiran Raja, and Christoph Busch. "Face Morphing Attack Generation & Detection: A Comprehensive Survey." In: *IEEE Transactions on Technology and Society* (2021), pp. 1–1.

[VEB09] Nguyen Xuan Vinh, Julien Epps, and James Bailey. "Information theoretic measures for clusterings comparison: is a correction for chance necessary?" In: *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*. 2009, p. 8.

[Wil18] Will Koehrsen. *Neural Network Embeddings Explained.* Oct. 2, 2018. URL: https://towardsdatascience.com/neural-network-embeddings-explained-4d028e6f0526 (visited on 11/30/2021).

[Zha+21] Haoyu Zhang, Sushma Venkatesh, Raghavendra Ramachandra, Kiran Raja, Naser Damer, and Christoph Busch. "MIPGAN - Generating Strong and High Quality Morphing Attacks Using Identity Prior Driven GAN." In: *IEEE Transactions on Biometrics, Behavior, and Identity Science* 3.3 (July 2021), pp. 365–383.

[Zha+16] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks." In: *IEEE Signal Processing Letters* 23.10 (Oct. 2016), pp. 1499–1503.