Identical Twins Face Morph Database Generation

Kelsey O'Haire, Sobhan Soleymani, Baaria Chaudhary, Jeremy Dawson, Nasser M. Nasrabadi
West Virginia University
{klo0003, ssoleyma, bac0062}@mix.wvu.edu, {jeremy.dawson, nasser.nasrabadi@mail.wvu.edu}

2022 IEEE International Joint Conference on Biometrics (IJCB) | 978-1-6654-6394-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/IJCB54206.2022.10007973

Abstract

By combining two or more face images of look-alikes, morphed face images are generated to fool Facial Recognition Systems (FRS) into falsely accepting multiple people, leading to failures in security systems. Despite several attempts in the literature, finding pairs of bona fide faces to generate the morphed images is still a challenging problem. In this paper, we morph identical twin pairs to generate extremely difficult morphs for FRS. We first explore three methods of morphed face generation, GAN-based, landmark-based, and a wavelet-based morphing approach. We leverage these methods to generate morphs from the identical twin pairs that retain high similarity to both subjects while resulting in minimal artifacts in the visual domain. To further improve the difficulty of recognizing morphed face images, we perform an ablation study to apply adversarial perturbation to the morphs such that they cannot be detected by trained morph classifiers. The evaluation of the generated identical twin morphed dataset is performed in terms of vulnerability analysis and presentation attack error rates.

1. Introduction

Facial recognition systems (FRS) are the most widely accepted method of biometrics at border security. The International Civil Aviation Commission (ICAO)'s [13] electronic Machine-Readable Travel Document (eMRTD) rely on facial recognition due to the ease of enrollment and cultural acceptance [3, 13]. Face data can also be verified by a human as needed, making it particularly attractive at border crossings where access to advanced verification technology may be limited. Indeed, even when FRS are considered, an image can be verified by a human as a last resort [4]. The four stages of a biometric system are enrollment, template creation, identification, and verification [13]. While FRS are a security necessity, they are vulnerable to attacks in the enrollment stage. If an enrolled passport photo resembles multiple people, the passport can be shared between the look-alikes.

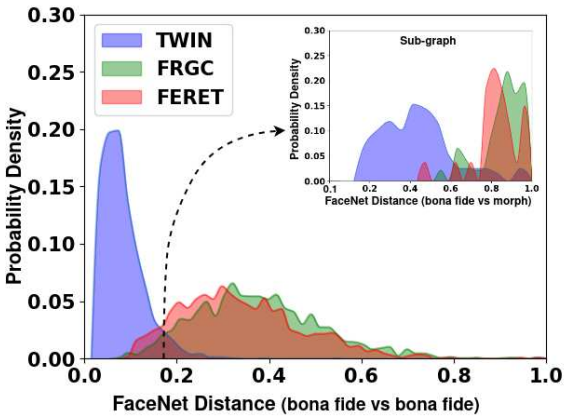

Figure 1: Probability density of the normalized FaceNet L_2 distances between the embeddings of bona fide subject pairings for the Twin, FRGC, and FERET datasets. At the distances of 0.17-0.21, the bona fide pairings are compared with their respective morphs. The sub-graph shows the probability density of the normalized FaceNet L_2 distances between the bona fide subjects and their respective morph.

Identical twins, also known as monozygotic twins, pose a severe problem to FRS since they represent the extreme scenario between individuals who naturally look alike [21]. Finding look-alikes is a necessary step when creating high-quality morphs in order to reduce artifacts and improve verification properties [5]. Paone *et al*. [21] studied a twins dataset made up of 126 twin pairs. They found that the Equal Error Rate (EER) for a twins dataset is significantly high as five of the seven algorithms tested had an EER at or above 50% for identical twins. Therefore, identical twins are a challenging paradigm for an FRS because of twins' inter-class similarity which can lead to high false acceptance rates in the verification stage [4, 21].

Morphed images are created by combining face images from two or more individuals, creating a new ambiguous face which possesses similarities between the bona fide identities. As morphing technology becomes more accessible, anyone can create high-quality morphed images with

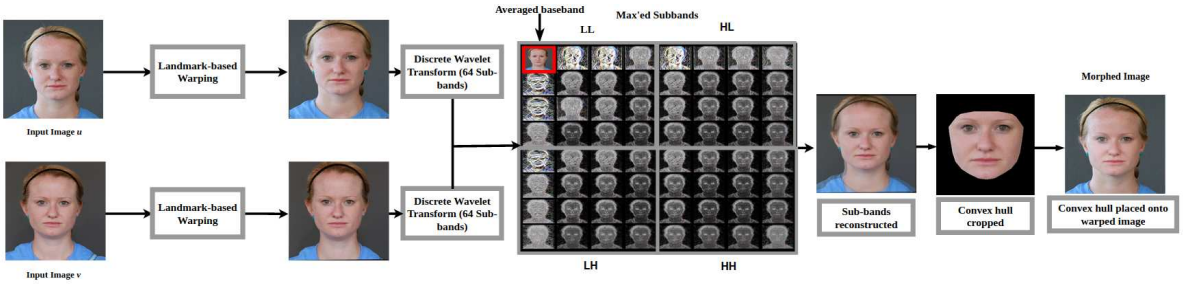

Figure 2: Wavelet-based morphing pipeline. The input subjects are warped, and then three-level wavelet decomposed into their respective 64 uniform sub-bands. Coefficient-wise, the lowest-frequency basebands are averaged together, and the remaining 63 sub-bands are maxed together. The resulting sub-bands are used to reconstruct the morphed image, cropped, and placed on the convex hull of the input subject.

little to no technical background, highlighting the need for more challenging morphed image datasets for training face morph detectors. Identical twins represent the ideal pairing condition for morphing and remove the ambiguity of finding look-alike pairs. The effectiveness of morphing twins is two-fold. First, Commercial Off-The-Shelf systems (COTS), as well as human verifiers, are vulnerable to the high-quality morphing attacks generated from similar face images [23]. Second, twins naturally looking similar creates ambiguity between individuals, causing an increase of false acceptance in detectors [4] and creating a very useful dataset for training and testing morph detectors.

To create an extremely hard scenario for an FRS, we generate a new dataset of identical twin morphed images. Our morphed faces are generated with three separate methodologies, landmark-based models [22, 18], Generative Adversarial Network (GAN)-based models [15, 14], and our wavelet-based morphing. Our Twin dataset provides ideal look-alike pairs for morphing. Consequently, we observe that the Twin dataset provides better morph generation capability compared to several other datasets across different morphing methodologies [24, 14]. As shown in Figure 1, for the same FaceNet [26] distance between the bona fide pairings, the twin morph dataset looks significantly more similar to its contributing bona fide identities than comparable morph datasets. In addition, for the bona fide pairs with similar distance, the twin dataset provides morphs harder to detect.[1]

2. Related Work

2.1. Morphing Techniques

Ferrara *et al*. morphed their images manually using the open-source image editor GIMP [9]. While the resulting images showed little artifacts, the pipeline was impractical to be scaled up for generation of large datasets. Since then, many open-source repositories have emerged, making it simple to generate large-scale datasets. Facemorpher [22], WebMorph [7], and OpenCV [18] are typical landmark-based algorithms that rely on a combination of warping and splicing to generate morphed images. While landmark-based morphing techniques are fast and effective, they tend to lead to warping artifacts in the high-frequency areas in the image, such as iris and outline of the face [29].

GAN-based morph generation creates morphs by combining the latent space representation of two face images. GANs have made strides in terms of quality and accessibility [14]. One of the most common GANs for morph generation in literature is StyleGAN2 [14, 15] because of its high-quality results and minimal artifacts. The networks are trained to generate high-quality reconstructions from a bottleneck stage. GAN-based morphing approaches have issues retaining identity information, causing morphs to be more heavily weighted toward one subject than another [24, 29]. MIPGAN-II [29] attempts to fix this problem by creating a loss-function based on perceptual-loss and identity priors in order to retain similarity to input subjects while creating high-quality morphs. This methodology was met with success, as the MIPGAN-II based morphs can fool multiple FRS at a higher rate than StyleGAN-based morphs.

2.2. Twin Face Morphing

Differentiating twins is a hard problem for facial recognition systems due to the high similarity between the two subjects [19]. In fact, even humans have a difficult time discerning between twin pairs. Biswas *et al*. [1] experimented with participants differentiating between twin pairs and images of the same person. They discovered that humans are only able to classify twin pairs versus images of the same individual at an average rate of 78.82%. There has been limited research on the effects of twins and facial recognition systems due to the lack of publicly available datasets. Paone *et al*. [21] studied a twins dataset made up of 126 twin pairs. They found that the Equal Error Rate (EER) for

[1]This database is available upon request: https://biic.wvu.edu/data-sets/identical-twin-face-morph-dataset

Table 1: MMPMR (%) at false match rate of 0.1%.

Dataset	Twin			FRGC [24]			
	Facemorpher	Wavelet	StyleGAN2	Facemorpher	OpenCV	StyleGAN2	MIPGAN-II
FaceNet	**97.70**	**98.11**	**93.01**	5.7	5.9	0.7	92.15
ArcFace	**99.20**	**99.41**	**94.54**	11.2	10.8	0.4	94.21

Dataset	FERET [24]			AMSL [20]		LMA-DRD [6]	
	Facemorpher	OpenCV	StyleGAN2	Facemorpher	StyleGAN2	Digital	Print+Scan
FaceNet	40.3	40.6	1.3	81.16	61.28	64.12	60.76
ArcFace	34.8	35.2	2.5	84.85	39.17	80.07	77.17

a twins dataset is significantly high. Five of the seven algorithms tested had an EER at or above 50% for identical twins.

Pair selection is a vital step when morphing faces, and can lead to drastic differences in quality of the morphed images [5, 25]. If the morph is to be passed between two individuals, the individuals must possess physical similarities. Scherhag *et al*. [25] proposed to classify a dataset into soft-biometrics such as hair color, skin color, age, and gender prior to morphing. Damer *et al*. [5] explored different methods of determining look-alikes and how they affect the quality of the morphs. They find a strong correlation between morphing similar looking individuals and higher attack rates. Morphing twin pairs represents the ideal scenario for morphing by removing the ambiguity of pairing look-alikes and guaranteeing high similarity between bona fide subjects.

3. Morphed Face Generation

In this section, we describe our morph generation algorithms. The landmark-based morphs are generated using the open source morphing library Facemorpher [22]. We utilize StyleGAN2 [14] as our GAN-based morphing model. Additionally, the wavelet-based landmark morphs are leveraged from undecimated wavelet decomposition for blending. For all three generated morph datasets, adversarial examples are generated to increase the difficulty of morph detection.

3.1. Landmark-based Morphing

Pair selection is crucial for landmark morphing. It is imperative that both subjects have similar facial structures to prevent extraneous morphing. Identical twins are ideal pairs for landmark-based morphing since they naturally have a similar face structure. With a high facial similarity, landmark points need to be warped a smaller distance resulting in less shadowing and higher-quality morphs. Additionally, twins have similar skin tones which leads to a more natural blending in the face morphing algorithms [24]. We utilize Facemorpher to generate our landmark morphed images

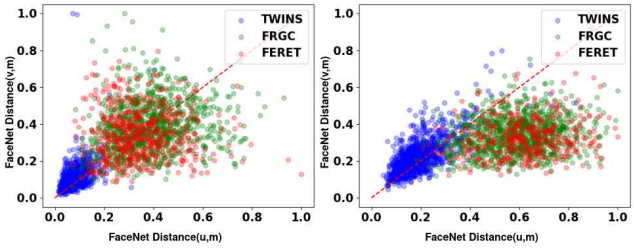

Figure 3: FaceNet L_2 distances between the bona fide faces and their respective morphs for the Twins, FRGC, and FERET datasets using (left) landmark and (right) StyleGAN2 morphing methods.

[22]. We consider two identical twin identities for morphing. Their respective face images $\boldsymbol{u}$ and $\boldsymbol{v}$ are to be aligned. We utilize a landmark-based approach where 68-landmark points are found on input images, creating the 68-element long pixel-coordinates $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{v}}$. Delaunay Triangles are utilized to create a mesh across the image, with the vertices of the mesh at $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{v}}$. The $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{v}}$ are averaged together to create the common landmarks coordinate, $\hat{\boldsymbol{m}}$. Bilinear interpolation is performed on the warped images to correct color values which results in the two face images sharing common landmark coordinates $\hat{\boldsymbol{m}}$. Then, an affine transform is used to map landmarked points from $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{v}}$ to the $\hat{\boldsymbol{m}}$ synthesizing $\hat{\boldsymbol{u}}_{\boldsymbol{w}}$ and $\hat{\boldsymbol{v}}_{\boldsymbol{w}}$. After warping, $\hat{\boldsymbol{u}}_{\boldsymbol{w}}$ and $\hat{\boldsymbol{v}}_{\boldsymbol{w}}$ are alpha-blended together in the following manner: $\alpha\hat{\boldsymbol{u}}_{\boldsymbol{w}} + (1-\alpha)\hat{\boldsymbol{v}}_{\boldsymbol{w}}$ to produce a blended image. We use alpha value of 0.5. Because of the alpha-blending, the background has heavy shadowing. Therefore, a convex hull is spliced from the blended image and placed back onto the face region of $\hat{\boldsymbol{u}}_{\boldsymbol{w}}$ to create the final morphed image, $\boldsymbol{m}$. Our algorithm is modified from Facemorpher at the stages where the background is warped and where the convex hull is spliced.

3.2. Wavelet Landmark Morphing

Our second method of landmark-based morphing leverages the spatial-frequency decomposition to fuse the

warped images. The twin pair images are aligned and warped in the same manner as the previous method. However, after the warping stage, $\hat{\boldsymbol{u}}_{\boldsymbol{w}}$ and $\hat{\boldsymbol{v}}_{\boldsymbol{w}}$ are decomposed into 64 equal sub-bands using a three-level undecimated wavelet decomposition. A vertical and a horizontal filter are applied to the warped images. We number the bands from $1, 2, \ldots, 64$, where the first sub-band represents the baseband. As presented in Figure 2, the lowest frequency sub-band (i.e. baseband) after three-level wavelet decomposition of the $\hat{\boldsymbol{u}}_{\boldsymbol{w}}$ and $\hat{\boldsymbol{v}}_{\boldsymbol{w}}$ are averaged together. This baseband is selected because it represents most of the shared information from the original subjects. The remaining 63 sub-bands are combined using the maximum-coefficient at every location in the sub-bands to capture the most significant information from each subject. If $[\boldsymbol{U}_1, \ldots, \boldsymbol{U}_{64}] = \Phi(\hat{\boldsymbol{u}}_{\boldsymbol{w}})$ and $[\boldsymbol{V}_1, \ldots, \boldsymbol{V}_{64}] = \Phi(\hat{\boldsymbol{v}}_{\boldsymbol{w}})$ are the undecimated wavelet decompositions of the aligned input faces, we define morphed sub-bands as:

$$\boldsymbol{M}_k[i,j] = \begin{cases} \text{mean}\left(\boldsymbol{U}_k[i,j], \boldsymbol{V}_k[i.j]\right), & k = 1 \\ \max\left(\boldsymbol{U}_k[i,j], \boldsymbol{V}_k[i.j]\right), & \text{otherwise.} \end{cases} \quad (1)$$

These morphed wavelet sub-bands are used to reconstruct the blended face image. The convex hull of the morphed image is spliced onto the background of $\hat{\boldsymbol{u}}_{\boldsymbol{w}}$ to create the morphed image. By averaging the wavelet coefficients from the lowest subbands, we capture most of the common information between the subjects. The Low-Low region of the images contains the general shape and color of subjects. This region is optimal for blending subjects' skin tone and general facial features. On the other hand, by maxing the wavelet coefficients in the high frequency subbands we capture the dominating features from each subject. These subbands contain the outline and structure of the face. Therefore, we max the coefficients in each sub-band for the aligned images to reinforce characteristics from each input image.

3.3. StyleGAN Morphing

We utilize the StyleGAN2 [14] to generate our morphed images. We begin with the same aligned twin pair images $\boldsymbol{u}$ and $\boldsymbol{v}$. First, they are warped using an affine transform to the same intermediate coordinates as outlined above to result in the warped images $\hat{\boldsymbol{u}}_{\boldsymbol{w}}$ and $\hat{\boldsymbol{v}}_{\boldsymbol{w}}$. The convex hulls of both images are spliced from the warped images and placed onto a black background. These convex hull images are embedded to 18×512 latent codes which are fused together to construct the morphed code. Custom noise is added to the convolutional layers to further texturize the morphs and increase the perceptual fidelity. This fused latent code is projected onto the generator to create the morphed convex hull as the convex hull between $\boldsymbol{u}$ and $\boldsymbol{v}$. This image is spliced back onto one of the input images to construct the morphed image $\boldsymbol{m}$.

3.4. Adversarially Perturbed Morph Generation

Goodfellow *et al*. [12] introduced the fast gradient sign method (FGSM), which perturbs the input of the model based on the sign of the gradient for a target class. Adversarial perturbations should not be perceptually visible in the adversarial images. Liao *et al*. [17] utilized FGSM with a masking technique to perturb areas deemed to possess high importance using the spatial information derived from multiple convolutional layers in a model. We add adversarial perturbation to the morphed images in order to further increase difficulty of their detection. FGSM perturbs an image based on the gradient with every iteration of backpropagation. Basic Iterative Method (BIM) [16] is a derivation of FGSM, where a constant step-size is utilized for every applied perturbation and an L_∞ constraint is used as maximum allowed pixel difference. Using a morph detector, images are perturbed:

$$\boldsymbol{m}_{N+1}^{adv} = \textit{Clip}_{m,\epsilon}\{\boldsymbol{m}_N^{adv} + \beta \text{sign}\left(\nabla_m L_{adv}\right)\}, \quad (2)$$

where $\boldsymbol{m}_0^{adv} = \boldsymbol{m}$ is the morph and L_{adv} consists of cross-entropy and Total Variation (TV) smoothing losses:

$$L_{adv} = J\left(\boldsymbol{m}_N^{adv}, y_{true}\right) - \lambda TV\left(\boldsymbol{m}_N^{adv}\right), \quad (3)$$

where J is the cross-entropy cost function between the adversarial image and the target class, β is the perturbation step size and ϵ is the L_∞ constraint on the pixel difference values [16]. The term y_{true} is equal to 1 for morphed images. The value of $\textit{Clip}_{m,\epsilon}$ confirms that the pixel values are within ϵ L_∞-norm distance from the original sample. We also clip the adversarial example at each iteration to make sure that all pixel values reside within the valid input range. Variable λ is the smoothness regularization parameter. To further improve the visual quality of the image, TV smoothing is applied to the perturbation image to remove any visible artifacts in the adversarial morphed image [27]:

$$TV(\boldsymbol{m}_N^{adv}) = \sum_{i,j} \Big((\boldsymbol{r}_N[i,j] - \boldsymbol{r}_N[i+1,j])^2 - (\boldsymbol{r}_N[i,j] - \boldsymbol{r}_N[i,j+1])^2 \Big)^{\frac{1}{2}}, \quad (4)$$

where $\boldsymbol{r}_N[i,j]$ is a pixel in the perturbation image $\boldsymbol{r}_N = \boldsymbol{m}_N^{adv} - \boldsymbol{m}$. We refer to the perturbed morph image as $\boldsymbol{m'}$. When perturbing an image, we use values $\beta = 6$, $\epsilon = 2$, and $\lambda = 0.55$ and perturb every morph image until the confidence score of the detector falls below 50%.

4. Experiments

We utilize our twin dataset to generate our morphs. The dataset contains images of sizes ranging from 2848×4288 to 5760×3840. Subjects have a neutral face and frontal pose angle, with images collected under controlled lighting

Table 2: Differential morph detection across datasets using FaceNet.

Dataset	Morph Type	AUC	APCER@BPCER			BPCER@APCER			EER
			1%	5%	10%	1%	5%	10%	
FRGC Facemorpher [24]	Landmark	99.82%	2.11%	0.63%	0.21%	6.55%	1.22%	0.616%	1.63%
FERET OpenCV [24]		99.00%	18.56%	4.92%	3.03%	14.77%	4.166%	4.16%	2.27%
FRGC OpenCV [24]		99.52%	8.40%	1.60%	0.8%	4.56%	1.52%	0.43%	2.61%
AMSL Facemorpher [20]		99.35%	18.41%	3.70%	1.01%	6.81%	3.31%	1.28%	3.68%
FERET Facemorpher [24]		98.91%	19.45%	5.83%	4.28%	16.97%	4.79%	1.10%	4.79%
LMA-DRD Print+Scan [6]		91.22%	75.68%	33.33%	29.63%	68.56%	41.86%	39.8%	16.27%
LMA-DRD Digital[6]		88.00%	80.26%	45.00%	27.50%	72.36%	57.50%	50.00%	20.00%
Twin Wavelet		**74.03%**	**61.48%**	**57.41%**	**51.51%**	**98.744%**	**85.33%**	**74.56%**	**33.71%**
Twin Landmark		**70.19%**	**64.73%**	**58.31%**	**54.09%**	**99.10%**	**90.20%**	**81.95%**	**36.84%**
FRGC StyleGAN2 [24]	GAN	99.65%	0.40%	0.12%	0.00%	0.42%	0.00%	0.00%	0.42%
AMSL StyleGAN2 [20]		99.98%	1.26%	0.00%	0.00%	0.86%	0.00%	0.00%	0.50%
MIPGAN-II [29]		99.85%	1.89%	0.27%	0.00%	5.06%	0.80%	0.26%	2.40%
FERET StyleGAN2 [24]		99.73%	4.29%	1.56%	0.00%	5.14%	1.56%	0.00%	2.94%
Twin StyleGAN2		**88.92%**	**41.83%**	**33.57%**	**27.77%**	**97.09%**	**62.25%**	**57.62%**	**19.95%**

with a neutral background. Every subject has a corresponding identical twin pair. There are a total of 2,268 unique identities in the twin dataset that generate a total of 2,978 morphed images per morphing method. Some subjects appear in the dataset more than once in different collections corresponding to different years. We refer to these morph datasets as Twin Landmark, Twin Wavelet, and Twin StyleGAN2. We use FaceNet [26] and ArcFace [8] as our verifiers. For the FaceNet verifier, we use MTCNN [30] to detect the faces and resize them to 160×160.

For the comparison of our landmark-based methods (landmark and wavelet morphs), we use seven different datasets found in literature. We use an FRGC and FERET morph dataset generated using Facemorpher and OpenCV created by Sarkar *et al*. [24], we refer to this dataset as FRGC Facemorpher, FRGC OpenCV, FERET Facemorpher, FERET OpenCV. Further, we utilize the LMA-DRD Digital and LMA-DRD Print+Scan datasets generated from the VGGFace-2 [2] dataset created by Damer *et al*. [6]. Lastly, we use the AMSL pairings of Neubert *et al*. [20] for our comparisons. To compare our StyleGAN2 generated morphs, we use four datasets found in literature. First, we use the FRGC and FERET morph datasets generated by Sarkar *et al*. [24]. We include a "state-of-the-art" FRGC dataset generated using MIPGAN-II [29]. Again, we use the AMSL pairings of Neubert [20] for our StyleGAN2 comparison. Each dataset has their own methodology for choosing the top look-alike pairs for morphing.

4.1. Vulnerability Analysis

To analyze the performance of the morphs, we utilize the International Organization for Standardization (ISO) [10] standards for reporting the performance, Attack Presentation Classification Error Rate (APCER) and Bona Fide Presentation Classification Error Rate (BPCER). The ISO describes APCER as the percentage of morphs incorrectly classified as bona fide presentations in a specific scenario. Inversely, BPCER is described as the percentage of the bona fide images incorrectly classified as presentation attacks [10]. We report APCER and BPCER values at the 1%, 5%, and 10%. In addition, we report the Area Under the Curve (AUC) and EER. We also use the Mated Morph Presentation Match Rate (MMPMR) as a metric to quantify the similarity between a generated morph image and its contributing subjects [25] where only morph-bona fide pairs with a similarity score above a threshold are considered:

$$\text{MMPMR}(\tau) = \frac{1}{M}\sum_{m=1}^{M}\left\{\left[\min_{n=1,..,N_m} S_m^n\right] > \tau\right\}, \quad (5)$$

where M is the total number of morphs and N_m is the number of subjects contributing to a particular morph [25]. S_m^n is the similarity score between the morph m and the n^{th} corresponding subject and τ is the operational verification threshold [25].

For our verifiers (i.e., FaceNet and ArcFace), we set the operational threshold at False Match Rate (FMR) of 0.1% [11]. Table 1 presents the MMPMR values for the Twin datasets compared to other baseline datasets. A higher MMPMR equates to a dataset with morphs that are more similar to their bona fide contributing subjects. For the Twin datasets, we observe that FaceNet consistently provides lower MMPMR% compared to ArcFace. By comparing the performance of the Twin datasets with the other morphs, we conclude that for both the landmark- and GAN-based datasets, the Twin morphed faces consistently excel. For instance, although MIPGAN-II benefits from a loss

Table 3: Differential morph detection across datasets using ArcFace.

Dataset	Morph Type	AUC	APCER@BPCER			BPCER@APCER			EER
			1%	5%	10%	1%	5%	10%	
AMSL Facemorpher [24]	Landmark	97.87%	16.25%	10.78%	6.29%	30.52%	16.82%	6.12%	8.11%
FRGC OpenCV [24]		96.60%	42.85%	15.81%	9.18%	53.18%	19.89%	9.85%	9.74%
FERET Facemorpher [24]		96.33%	26.46%	14.55%	10.77%	57.46%	26.55%	12.66%	10.68%
FRGC Facemorpher [20]		96.85%	27.55%	15.85%	11.22%	53.06%	17.34%	10.87%	10.71%
FERET OpenCV [24]		96.32%	24.16%	14.93%	10.77%	57.18%	27.88%	12.47%	10.77%
LMA-DRD Print+Scan [6]		90.26%	67.78%	39.54%	31.97%	53.48%	43.64%	30.32%	17.35%
LMA-DRD Digital[6]		88.88%	65.25%	40.75%	29.63%	57.12%	39.31%	35.70%	21.25%
Twin Landmark		**71.01%**	**81.36%**	**67.44%**	**57.76%**	**98.44%**	**91.77%**	**84.77%**	**34.36%**
Twin Wavelet		**69.98%**	**84.56%**	**71.61%**	**59.93%**	**99.52%**	**92.38%**	**85.27%**	**34.49%**
AMSL StyleGAN2 [20]	GAN	99.96%	0.07%	0.00%	0.00%	0.00%	0.00%	0.00%	0.97%
FRGC StyleGAN2 [24]		99.85%	0.18%	0.06%	0.00%	0.00%	0.00%	0.00%	1.03%
FERET StyleGAN2 [24]		99.75%	3.59%	0.95%	0.37%	10.20%	1.03%	0.05%	2.55%
MIPGAN-II [29]		99.07%	10.65%	6.35%	2.18%	20.25%	8.07%	1.91%	5.91%
Twin StyleGAN2		**93.60%**	**28.39%**	**19.90%**	**15.38%**	**89.67%**	**45.67%**	**22.33%**	**13.58%**

function to exploit perceptual quality and identity factor, but the Twin StyleGAN2 dataset still provides better performance. Additionally, we see evidence that the Wavelet dataset retains identity better than any of the other datasets which can be attributed to the wavelet coefficient maxing where we retain the dominating features of the face.

4.2. Dataset Comparison

For an initial observation of the quality of our morphs, we plot the normalized L_2 FaceNet distance between bona fide pairings found in Figure 1. FaceNet is trained in such a way that smaller L_2 distance between embeddings equates to a stronger look-alike pair [26]. We use the FaceNet distances between the bona fide subject pairings for the Twin, FERET and FRGC morph [24] datasets. The pairs of the FERET and FRGC datasets are generated lexographically, where pairs of the same gender and similar demographics are paired together. For our FaceNet verifier, we set the FAR=10^{-3} and find the statistical equivalent of the bona fide to bona fide distance that results in morphs accepted in the differential setting. Thus, we consider the normalized distance below 0.21 to be a strong look-alike which will create high-quality morphed images. It is obvious from Figure 1 that the Twin pairings have an advantage over the other compared datasets. Of 1,134 pairs, the Twins dataset contains 1,105 pairs below this threshold (97.4% of pairs) compared to 82 of the 964 (8.50% of pairs) in FRGC and 80 of the 529 (15.1% of pairs) in FERET. In order for the FRGC dataset to have the same number of high-quality pairs below the 0.21 threshold as the Twin dataset, it would require the number of paired individuals to increase from 964 to about 11,500 or in other words, about 23,000 subjects needed in the dataset. Generating a datasets of about 23,000 subjects with passport-quality images requires tremendous time and effort and would be near impossible for most datasets that would be made available in literature, which typically have around 1,200 subjects [24, 20, 5].

In Figure 1, a threshold is placed on the distribution to capture the pairings between the distances of 0.17-0.21. This is an arbitrary threshold centered around the peak intersection of similarity between datasets at distance 0.19. This is an area of high interest because it is the area of maximum overlap between datasets and most applicable for direct comparison. Using the selected look-alike pairings in this region , we find the normalized FaceNet distance between the pair of subjects and their respective morphs as seen in the sub-graph of Figure 1. Clearly, even with the same similarly measure between bona fide subjects , there is a correlation between the quality of pairs and the resultant morph's ability to retain identity with its contributing subjects. Because the twin dataset has stronger pairings, the generated morphs look more like their bona fide subjects after morphing giving them stronger attack abilities.

To further understand the relationship between the morphs and their contributing subjects, we plot the FaceNet distance scores between all of the morphs and with their bona fide subjects in Figure 3. The x-axis represents the FaceNet L_2 distance between subject 1 and the morph, where the y-axis represents subject 2 to the morph. We compare the FaceNet distances of the Twins database to the FRGC and FERET morphs. Again, we observe that the twin morphs consistently having lower FaceNet distances than the FRGC and FERET datasets. Further, we observe that the twin dataset has a lower variance in distance scores for both the landmark and StyleGAN2 settings. Meaning that not only do the twin morphs show high similarity,

Table 4: Universal and dedicated FaceNet morph detectors tested on morph and perturbed morph Twin images.

	Dataset	AUC	APCER@BPCER			BPCER@APCER			EER
			1%	5%	10%	1%	5%	10%	
Universal	Landmark	63.62%	96.86%	90.93%	80.81%	93.38%	86.61%	76.69%	40.56%
	Landmark Perturbed	56.22%	98.48%	93.83%	85.23%	95.80%	92.17%	84.67%	45.56%
	StyleGAN2	78.79%	95.23%	74.18%	53.13%	79.77%	69.55%	58.22%	28.66%
	StyleGAN2 Perturbed	71.88%	94.65%	80.46%	69.65%	86.66%	77.33%	68.22%	34.26%
	Wavelet	70.55%	96.61%	87.56%	75.12%	87.50%	78.67%	67.40%	34.55%
	Wavelet Perturbed	62.75%	97.94%	91.78%	81.15%	92.57%	85.56%	77.52%	39.79%
Dedicated	Landmark	80.28%	90.00%	62.55%	47.79%	86.12%	70.48%	58.46%	26.53%
	Landmark Perturbed	50.41%	98.37%	91.27%	85.00%	99.43%	97.17%	94.67%	49.35%
	StyleGAN2	92.20%	90.00%	36.27%	21.86%	53.55%	32.00%	21.11%	14.66%
	StyleGAN2 Perturbed	82.77%	94.53%	65.11%	48.02%	78.22%	57.33%	44.00%	25.77%
	Wavelet	78.63%	88.04%	74.63%	54.95%	87.01%	68.38%	50.73%	28.67%
	Wavelet Perturbed	59.10%	99.26%	95.83%	92.64%	90.77%	84.06%	75.00%	43.52%

they retain identity significantly better than the FRGC and FERET morphs. This anomaly is especially apparent in the StyleGAN2 datasets, where the FRGC and FERET datasets clearly bias toward one subject.

4.3. Morph Detection

We analyze the twin morphs in a differential scenario and compare them to several other datasets. Historically, morph datasets do not perform well in a verification scenario [29, 24] since FRS are exceptional at differentiating between a bona fide and false image of a target individual. This scenario becomes very difficult when the morph image is extremely close to the target individual as is the case when morphing identical twins. We consider the L_2 distance measure between the embeddings of FaceNet [26] and ArcFace [8] to detect morphed images in a differential morph scheme. The results from the differential morph detection can be found in Tables 2 and 3. Using the distance score from the respective verifiers, we compare the verification capability of our twins datasets to those morph datasets found in literature. For the FaceNet verifier, the twins datasets are able to achieve an AUC of 74.03%, 70.19%, and 88.92% for the Twin Wavelet, Twin Landmark, and Twin StyleGAN2 datasets, respectively. All three twin datasets have an EER above 18%, meaning that the morphs are highly effective at being verified as their bona fides.

For the landmark-based datasets in a differential morph detection setting, our twin morphs preform significantly better than the comparison datasets, with our Twin Landmark morphs achieving an EER of 36.84% and 34.36% on FaceNet and ArcFace, respectively. When comparing to the rest of the datasets, all datasets have EER values below 21% across both FaceNet and ArcFace. Using FaceNet, five of the seven datasets (FRGC Facemorpher, FERET OpenCV, AMSL Facemorpher, FERET Facemorpher) have EER values below 5%. Both the Twin Landmark and Twin Wavelet morphs have an AUC of approximately 14% lower than the best performing landmark comparison dataset using FaceNet (LMA-DRD Digital with AUC of 88.00%). We see that the Twins Wavelet morphs outperform all of the compared Landmark datasets, this is because of the maxing step of the wavelet morphing which may be helping retain identity in the morph. In a differential scenario, it is clear that the twin morphs retain the identity of the bona fide subjects better than the compared datasets.

Comparing our StyleGAN2 morphs, the Twin dataset performs with an AUC of 88.92% with FaceNet and 93.00% using ArcFace. The Twin StyleGAN2 images result in the highest EER value across both verifiers. In this scenario, we observe the issue GANs have with retaining identity information, with the EER of all GAN-generated morphs significantly lower than their respective landmark datasets. For instance, the Twin Landmark dataset, using FaceNet as a verifier, has an EER value of 33.71% while the StyleGAN2 generated data has an EER value of 19.95%. This pattern can also be observed in the FRGC Facemorpher and FRGC StyleGAN2 dataset, having an EER of 1.63% and 0.42%, respectively. On the Twin datasets, ArcFace is better equipped to differentiate between the morphed image and bona fide subject. However, ArcFace underperformed with the Twin Wavelet, compared to Twin landmark.

5. Ablation Study

While our datasets are already a challenge for FRS, we further improve the attack capability of the morph images by adversially perturbing them. By perturbing the images, our goal is to fool morph detectors into classifying our morph images as genuine. For perturbing, we need two separate groups of detectors, one group to perturb images and then a second group act as independent morph detec-

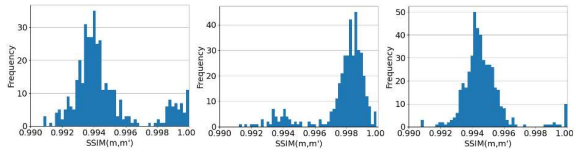

Figure 4: SSIM score distribution between the original and perturbed morphs. The distributions show the Landmark, Wavelet, and StyleGAN2 datasets from left to right.

tors to test the perturbation. Using the Inception-Resnet v1 [28] pretrained on VGGFace2 [2], we fine-tune four morph detectors. Our universal perturbing detector is trained on all three morph datasets, while the dedicated perturbing detectors are trained on each of our three morph datasets to detect morph imagery. We refer to these models as the universal perturbing and dedicated perturbing morph detectors. Prior to perturbation, the Landmark, StyleGAN2, and Wavelet datasets have a classification AUC values of 95.96%, 99.83%, and 99.80% on the dedicated perturbing detectors, respectively. After perturbation, the AUC value of all datasets drops significantly to 46.98%, 56.84%, and 27.87%, respectively for the dedicated perturbed datasets. The universal perturbing detector is used to perturb the datasets as well, seeing AUC values of 99.30% perturbed to 80.53% for StyleGAN2, 97.90% to 57.73% for Landmark, and 99.44% to 55.06% for Wavelet to create the universal perturbed dataset. We use Structural Similarity Index Measure (SSIM) as a metric of determining image quality of the perturbed images [30]. We find the SSIM score between the original image as a reference and the perturbed images from the universally perturbed dataset. A higher SSIM score means a higher similarity between the perturbed and unperturbed images. All morphs after perturbation have an SSIM score above 0.99 with their original morph counterparts which illustrates that the perturbation applied to the morphs is imperceivable as shown in Figure 4. As shown in Figure 5, the perturbed morphs maintain their visual quality.

To test the efficacy of the perturbation on a independent detectors, we extract FaceNet features and train a two-node binary classifier to classify images as bona fide or morph. We train four models, in the same manner we trained the perturbing detectors. We call these the universal FaceNet detector and the dedicated FaceNet detectors. A testing subset of the universal perturbed datasets is tested on the universal FaceNet detector and the results are presented in Table 4. Additionally, the dedicated datasets are tested using the FaceNet dedicated detectors. We observe an increase in EER and in APCER@BPCER=5% across all datasets. When ACPER@BPCER increases, the morphs are being labeled as bona fide at a higher rate. This shows that the quality of the morphing attack is improved after adding perturbation and that the pertubation affects the accuracy across

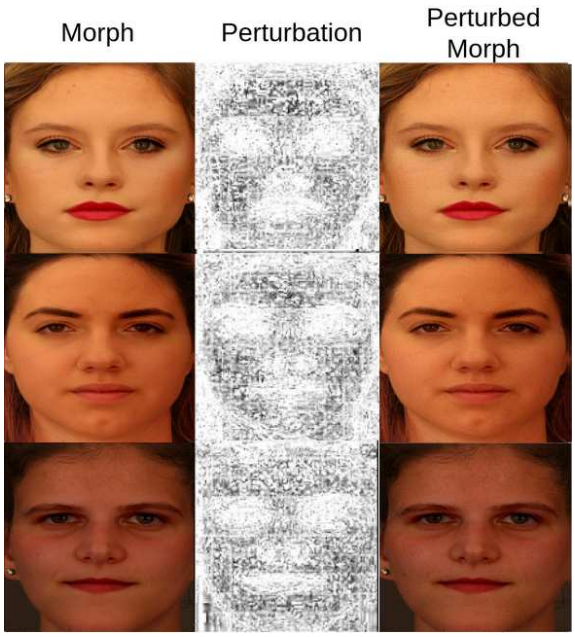

Figure 5: The left and right columns shows the twin morph and perturbed morph, respectively. The middle column shows the normalized perturbation applied to the morphed image.

"unseen" detectors. This transferability shows that our perturbed Twin dataset is even more difficult to detect.

6. Conclusion

In this paper, we generated morphed faces from identical twins with strong morphing attack capabilities. Further, we showed morphing using an undecimated wavelet decomposition can lead to stronger morphs. Morphing twins is a significant challenge for FRS that leads to erroneous verification, with our twin datasets scoring over 10% AUC lower than datasets found in literature. We showed that the twin morphs represent an extremely difficult scenario for FaceNet, leading to abnormally high error rates. With FaceNet EER values above 30% for all three twin datasets, the need for more work on these extreme cases is emphasized. To further improve the attack quality of our morphs, we explored the effect of adding adversarial perturbation to our morphs and showed that the perturbation is transferable across several unseen classifiers. The perturbation gave the already difficult twin morph dataset even greater capabilities. The generated twin morphs are one of the ultimate challenges for an FRS and can be used to further test the accuracy of morph detectors.

ACKNOWLEDGEMENT

This work is based upon a work supported by the Center for Identification Technology Research and the National Science Foundation under Grant #1650474.

References

[1] S. Biswas, K. Bowyer, and P. J. Flynn. A study of face recognition of identical twins by humans. *IEEE International Workshop on Information Forensics and Security*, pages 1–6, 2011.

[2] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *IEEE international conference on automatic face & gesture recognition (FG)*, pages 67–74, 2018.

[3] L. R. Carlos-Roca, I. H. Torres, and C. F. Tena. Facial recognition application for border control. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1–7, 2018.

[4] B. Chaudhary, P. Aghdaie, S. Soleymani, J. Dawson, and N. M. Nasrabadi. Differential morph face detection using discriminative wavelet sub-bands. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1425–1434, 2021.

[5] N. Damer, A. M. Saladie, S. Zienert, Y. Wainakh, P. Terhörst, F. Kirchbuchner, and A. Kuijper. To detect or not to detect: The right faces to morph. In *International Conference on Biometrics (ICB)*, pages 1–8, 2019.

[6] N. Damer, N. Spiller, M. Fang, F. Boutros, F. Kirchbuchner, and A. Kuijper. PW-MAD: Pixel-wise supervision for generalized face morphing attack detection. In *International Symposium on Visual Computing*, pages 291–304, 2021.

[7] L. DeBruine. debruine/webmorpher: Beta release 2, January 2018.

[8] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

[9] M. Ferrara, A. Franco, and D. Maltoni. The magic passport. In *IEEE International Joint Conference on Biometrics*, pages 1–7, 2014.

[10] I. O. for Standardization. ISO/IEC DIS 30107-3:2016: Information technology biometric presentation attack detection, 2017.

[11] FRONTEX. Best practice technical guidelines for automated border control (ABC) systems, 2015.

[12] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *Conference on Learning Representations*, 2015.

[13] ICAO. 9303-machine readable travel documents-part 9: Deployment of biometric identification and electronic storage of data in eMRTDs. *International Civil Aviation Organization (ICAO)*, 2015.

[14] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[15] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of StyleGAN. In *IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

[16] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

[17] Q. Liao, Y. Li, X. Wang, B. Kong, B. Zhu, S. Lyu, Y. Yin, Q. Song, and X. Wu. Imperceptible adversarial examples for fake image detection. In *IEEE International Conference on Image Processing (ICIP)*, pages 3912–3916, 2021.

[18] S. Mallick. Face morph using opencv - c++/python, March 2016.

[19] J. McCauley, S. Soleymani, B. Williams, J. Dando, N. Nasrabadi, and J. Dawson. Identical twins as a facial similarity benchmark for human facial recognition. In *International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5, 2021.

[20] T. Neubert, A. Makrushin, M. Hildebrandt, C. Kraetzer, and J. Dittmann. Extended StirTrace benchmarking of biometric and forensic qualities of morphed face images. *IET Biometrics*, 7(4):325–332, 2018.

[21] J. R. Paone, P. J. Flynn, P. J. Philips, K. W. Bowyer, R. W. V. Bruegge, P. J. Grother, G. W. Quinn, M. T. Pruitt, and J. M. Grant. Double trouble: Differentiating identical twins by face recognition. *IEEE Transactions on Information forensics and Security*, 9(2):285–295, 2014.

[22] A. Quek. Facemorpher, Jan 2019.

[23] D. J. Robertson, R. S. Kramer, and A. M. Burton. Fraudulent id using face morphs: Experiments on human and automatic recognition. *PLoS One*, 12(3):e0173319, 2017.

[24] E. Sarkar, P. Korshunov, L. Colbois, and S. Marcel. Vulnerability analysis of face morphing attacks from landmarks and generative adversarial networks. *arXiv preprint arXiv:2012.05344*, 2020.

[25] U. Scherhag, A. Nautsch, C. Rathgeb, M. Gomez-Barrero, R. N. Veldhuis, L. Spreeuwers, M. Schils, D. Maltoni, P. Grother, S. Marcel, et al. Biometric systems under morphing attacks: Assessment of morphing techniques and vulnerability reporting. In *International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–7, 2017.

[26] F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.

[27] M. Sharif, S. Bhagavatula, L. Bauer, and M. Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. *ACM SIGSAC Conference on Computer and Communications Security*, 2016.

[28] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, Inception-ResNet and the impact of residual connections on learning. In *AAAI Conference on Artificial Intelligence*, 2017.

[29] H. Zhang, S. Venkatesh, R. Ramachandra, K. Raja, N. Damer, and C. Busch. MIPGAN-Generating strong and high quality morphing attacks using identity prior driven GAN. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(3):365–383, 2021.

[30] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.