Juvenile Morph Dataset: A Study of Attack Detectability and Recognition Vulnerability

Kelsey O'Haire*, Sobhan Soleymani*, Samuel Price, Nasser M. Nasrabadi
Lane Department of Computer Science and Electrical Engineering
West Virginia University
Morgantown, West Virginia, USA
{klo0003, ssoleyma, swp0001}@mix.wvu.edu, nasser.nasrabadi@mail.wvu.edu

2022 IEEE International Symposium on Technologies for Homeland Security (HST) | 978-1-6654-9404-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/HST56032.2022.10025448

***Abstract*—A morph is an image of an ambiguous subject generated by combining multiple individuals. The morphed image can be submitted to a facial recognition system and erroneously verified with the contributing bad actors. When submitted as a passport image, a morphed face poses a national security threat because a passport can then be shared between the individuals. As morphed images become easier to generate, it is vital that the research community expands available datasets in order to contentiously improve current technology. Children are a challenging paradigm for facial recognition systems and morphing children takes advantage of this disparity. In this paper, we morph juvenile faces in order to create a unique, high-quality dataset to challenge FRS. To the best of our knowledge, this is the first study on the generation and evaluation of juvenile morphed faces. The evaluation of the generated morphed juvenile dataset is performed in terms of vulnerability analysis and presentation attack error rates.**

***Index Terms*—Juvenile Morphing, GAN-based Morphing, Landmark Morphing**

I. Introduction

Public acceptance and easy enrollment process make the face the most readily accessible form of biometric. Additionally, face images are relatively easy for humans to verify in-person without the need for sophisticated verification technology, making it attractive to border security. Therefore, the International Civil Aviation Commission (ICAO) [1] utilizes facial recognition for it's electronic Machine-Readable Travel Document (eMRTD) [1]. The reliability of facial recognition systems (FRS) is threatened by false positives, which allows an individual to be erroneously verified by the system as a different individual. These false positives can occur when subjects look alike and the FRS is not precise enough to differentiate between the individuals. Morphed images take advantage of this vulnerability. Morphed faces are generated by combining look-alike individuals in to an ambiguous face image which is verified as both individuals [2]. Morphed images of look-alikes are effective at fooling FRS, which poses a significant security threat [2], [3]. If a bad actor submits a morphed image to a passport enrollment system, the passport may be shared between multiple individuals.

* Authors Contributed Equally.

978-1-6654-9404-5/22/$31.00 ©2022 IEEE

Facial recognition systems perform lower on children than adults. Michalski *et al.* show that commercial-off-the-shelf (COTS) algorithms at an FMR of 0.1% in a verification setting for juveniles result in a false match rate up to six times higher than adults [4]. One of the major barriers to the improvement of juvenile face recognition is the lack of publicly available datasets dedicated to children [4]. Most FRS common in literature are trained on large publically-available datasets such as Visual Geometry Group Face2 (VGGFace2) [5]. While these datasets contain children's faces, the proportion of juvenile subjects is statistically insignificant to create reliable FRS for children. Srinivas *et al.* [6] study multiple COTS and government-off-the-shelf (GOTS) algorithms to understand the bias FRS have against children. They were able to deduce that in both identification and verification scenarios, children do not perform as well as adult baselines. Similarly, the Face Recognition Vendor Test (FRVT) [4], [7] has consistently shown lower performance on child subjects than on adult faces.

Additionally, children are more difficult to verify in person than adult subjects, creating a challenge for in-person verification which would otherwise come naturally [8]. We propound this crucial scenario with serious implications for national security and child trafficking: If a bad actor attempts to cross an international border with a child, the bad actor can create a morphed image of the child with a look-alike and pass the child through border security under the doppelganger's alias. In 2019, in the United States alone, there were over 6,000 reported cases of adults crossing a border with a minor fraudulently labeled as their own [9]. Our work is vital to detecting vulnerable children in these scenarios.

To the best of our knowledge, this is the first attempt to morph juvenile subjects to create morphed faces. We generate and evaluate 52,686 high-quality morph images utilizing two landmark-based and one generative adversarial morph method for children of the wide age range of 4 to 17 years old. Examples of our generated morphs from each of our morphing techniques can be found in Fig. I. These images present a difficult scenario for face verification systems and can be utilized to improve FRS models, as well as shed light on the current dangers of morphing children's faces. Many deep learning models show a strong bias against children [7], by morphing children we take advantage of this bias in order to further fool facial recognition systems.

Fig. 1. The bona fide subjects and morphed samples from UNCW dataset.

TABLE I
MORPH DETECTION PERFORMANCE ON OUR SIX MORPHED DATASETS.

	Morph Dataset	AUC	APCER@BPCER			BPCER@APCER			EER
			1%	5%	10%	1%	5%	10%	
Differential	Clarkson StyleGAN2	89.75%	46.55%	36.72%	28.85%	72.86%	55.85%	32.44%	16.73%
	Clarkson OpenCV	83.58%	58.57%	51.13%	44.33%	76.32%	67.45%	48.81%	24.74%
	Clarkson Facemorpher	83.86%	54.85%	47.76%	40.29%	75.68%	71.13%	52.97%	24.70%
	UNCW StyleGAN2	97.32%	27.22%	13.70%	5.05%	42.40%	15.00%	8.51%	9.44%
	UNCW OpenCV	92.23%	48.15%	28.97%	18.03%	77.66%	44.97%	30.17%	14.64%
	UNCW Facemorpher	89.45%	53.90%	40.75%	30.78%	80.52%	51.45%	37.20%	18.45%
Single	Clarkson StyleGAN2	79.68%	86.57%	69.57%	52.47%	97.91%	71.00%	55.39%	27.06%
	Clarkson OpenCV	69.89%	93.15%	75.29%	65.60%	99.52%	91.77%	78.48%	36.12%
	Clarkson Facemorpher	70.47%	94.52%	74.33%	64.78%	97.76%	92.00%	80.56%	33.54%
	UNCW StyleGAN2	92.66%	72.35%	29.39%	21.44%	71.15%	40.94%	26.17%	14.55%
	UNCW OpenCV	81.38%	95.34%	76.02%	58.12%	74.07%	59.39%	44.24%	24.77%
	UNCW Facemorpher	81.11%	95.59%	73.32%	58.88%	73.25%	60.01%	45.17%	27.29%

II. JUVENILE MORPHED FACE GENERATION

Here, we utilize our modified Facemorpher [10], OpenCV [11], and StyleGAN2 [12] to generate high-quality morphs.

A. Landmark-based Morphing

Landmark-based morphed image generation typically consists of three steps: landmark detection, warping, and blending. The landmark points of the two input subjects, which are critical points on each face, are averaged together to create a common set of landmarks. The images are then warped towards these common landmarks and blended to create the morphed image. The morphed images are guaranteed to have visual similarity with both individuals because features of the individuals are combined by averaging the input images together. Ferrara *et al.* [2] was the first to expose the dangers of morphed images in FRS by morphing high-quality images by hand. Sarkar *et al.* [3] generated data from various landmark-based algorithms such as Facemorpher [10] and OpenCV [11].

We consider two look-alike individuals for morphing. The pair's faces, $\boldsymbol{u}$ and $\boldsymbol{v}$, are aligned. 68-element long pixel-coordinates $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{v}}$ are found on each subject's face. The landmark coordinates are areas deemed of high importance for morphing. Then, $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{v}}$ are used to generate a mesh grid across the image. On an element-wise basis, the coordinates of $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{v}}$ are averaged together to create the common landmarks coordinate, $\hat{\boldsymbol{m}}$. After warping to the common landmarks, bilinear interpolation is performed in order to correct color values. An affine transform is used to transmute points from $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{v}}$ to the $\hat{\boldsymbol{m}}$ creating both $\hat{\boldsymbol{u}}_{\boldsymbol{w}}$ and $\hat{\boldsymbol{v}}_{\boldsymbol{w}}$. After warping, $\hat{\boldsymbol{u}}_{\boldsymbol{w}}$ and $\hat{\boldsymbol{v}}_{\boldsymbol{w}}$ are averaged together. At this point, the background of the face regions will have a heavy ghosting effect. Face region is spliced from the background and placed onto the convex hull of $\hat{\boldsymbol{u}}_{\boldsymbol{w}}$ to generate the final image $\boldsymbol{m}$. Our algorithm is modified from both Facemorpher [10] and OpenCV [11] at the stages where the background is warped and where the convex hull is spliced.

B. StyleGAN2 Morphing

In recent years, Generative Adversarial Networks (GANs) have become more powerful, by creating realistic looking images with minimal visual artifacts [12]. GAN-based morph generation approaches use latent vectors of input images which are then linearly combined, resulting in minimal artifacts and producing high-quality morphs [3], [12]. Damer *et al.* introduced MorGAN [13] for face morphing. They utilize their discriminator and generator in order to learn the mappings for the encoder and decoder. The networks are trained to generate reconstructions from the information bottleneck. Once MorGAN was trained, the latent vectors were linearly combined to generate the morphed image.

We combine the latent code using StyleGAN2 [12] to generate our morphed images because of the high-visual quality of their output images. While GAN-bsed approaches are becoming more popular, literature shows that GAN-generated

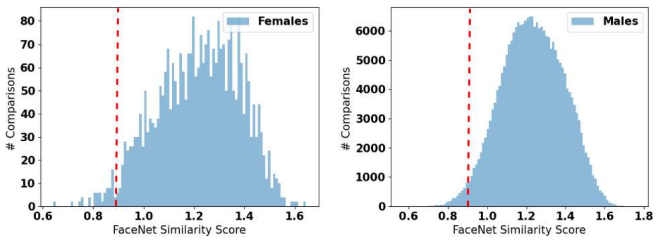

Fig. 2. All-to-all distribution for comparisons of subjects from the UNCW dataset. Pairs below the distance threshold are considered look-alikes.

morphs struggle to retain the identity of the input subjects [3]. The same aligned pairs as described in the previous section are used as $\boldsymbol{u}$ and $\boldsymbol{v}$. They are warped toward common landmarks in the same manner to result in the warped faces $\hat{\boldsymbol{u}}_{\boldsymbol{w}}$ and $\hat{\boldsymbol{v}}_{\boldsymbol{w}}$. The face region of both warped images are spliced and pasted onto a black background. These images are embedded to an 18×512 latent code. These codes are then averaged together to construct the morphed image's latent code. To improve final visual quality of the morphs, custom noise is added to the convolutional layers of StyleGAN2. This fused latent vector is reconstructed to generate the morphed convex hull. This face image is spliced back onto the face region of the input images $\boldsymbol{u}$ to construct the morphed image $\boldsymbol{m}$.

III. Experiments and Discussions

Two datasets are utilized to create our morphed images, the Clarkson University children dataset [14] and UNCW MORPH age-progression dataset [15]. The two datasets are utilized in order to generate a range of generated ages, with the Clarkson dataset containing images of children ages 4-11 years old, and a subset of the UNCW dataset containing subjects ranging from 16-17 years old. For both datasets, the subjects are in front of a neutral background and looking directly into the camera. A four year old has vastly different facial features than a 17 year old. When morphing, it is vital to morph subjects who look-alike in order to reduce morphing artifacts. Therefore, we preserve the integrity of the demographics of each dataset by generating two separate morphed datasets from the respective bona fide datasets.

UNCW dataset: From [15], we extract individuals of age 16-17 years old. The dataset has a strong gender bias, and our subset includes 499 male and 58 female subjects. The images are of size 470×400. Compared to Clarkson dataset, the subjects in this dataset have highly distinguishable features, similar to adults. We use the L_2 distance between the FaceNet's embeddings of length 512 in order to generate a similarity scores [16]. Morphs are generated within gender groups, and similarity scores are calculated within genders. As presented in Fig. 2, distance threshold is set at the top 5% of the female pairs in the distribution and pairs below this threshold are considered look-alikes. This threshold is also applied to the male distribution. 465 subjects are accounted in the final pairings, and per morphing method, 7,564 morphs are generated. We refer to the generated juvenile UNCW morph datasets using Facemorpher [10], OpenCV [11], and StyleGAN2 [12] as UNCW Facemorpher, UNCW OpenCV, and UNCW StyleGAN2, respectively.

Clarkson dataset [14] is made up of children ages 4-11 years old. The original images are of sizes 5472×3648 and of good visual quality. We used a subset of the data containing 165 subjects. The children are so young their faces lack highly distinguishable features, thus, creating high inter-class similarity between the subjects. Therefore, using FaceNet we find the top 10,000 look-alike pairs and use them for morphing. We crop the images to 512×512 and morph using the Facemorpher landmark-based, OpenCV landmark-based, and StyleGAN2 techniques. The resulting images are 512×512 and have no visual morphing artifacts. We refer to these three datasets as Clarkson Facemorpher, Clarkson OpenCV, and Clarkson StyleGAN2, respectively.

A. Vulnerability Analysis

Morphed images contain structural similarities with their bona fide subjects. Structural Similarity Index Measure (SSIM) [17] is calculated based on perceived similarity between reference images rather than a pixel-to-pixel comparison. As presented in Fig. III-A, we compare the SSIM score between the bona fide identities and their respective morphs. A higher SSIM score represents greater structural similarity. The datasets show a linear correlation between the structural similarities of bona fide identities and the morphed image. While the Clarkson dataset's SSIM scores trend higher than UNCW, it has a higher variance. Meaning, the Clarkson dataset maintained similarity better than UNCW, but shows greater bias toward one contributing subject over another. This is due to the greater variable face shapes in the young children in the Clarkson dataset. Therefore, when the convex hull is placed onto a contributing subject's face the SSIM is biased toward the subject used as the background of the morphed face, *i.e.*, the image with stronger structural similarities.

The International Organization for Standardization (ISO) defines Attack Presentation Classification Error Rate (APCER) as the rate of incorrectly identified morphed images while Bona Fide Presentation Classification Error Rate (BPCER) is the number of bona fide images erroneously labeled as a morph [18]. Further, we include the Mated Morph Presentation Match Rate (MMPMR) as a means of similarity between our morph images and its contributing subjects [19] where only morph/bona fide pairs which have a similarity score above a given threshold are considered:

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

where M is the number of morph images, N_m is the number of subjects contributing to a given morph, S_m^n is the similarity score between the morph $\boldsymbol{m}$ and its n^{th} corresponding subject [19]. As presented in Table II, we use FaceNet [16] and ArcFace [20] as our verifiers with τ as the operational verification threshold at False Match Rate (FMR) of 0.1% [21].

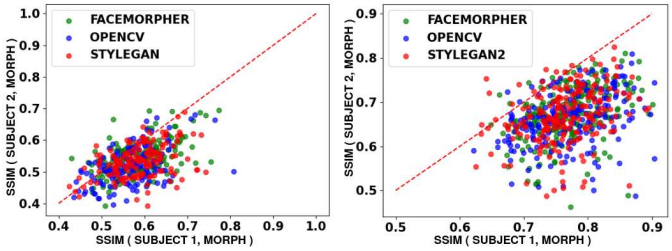

Fig. 3. SSIM between scores between bona fide and morphed images for the UNCW (left) and Clarkson (right) datasets.

For our six juvenile morph datasets, FaceNet is more vulnerable compared to ArcFace. In addition, the landmark morphing datasets provide higher vulnerability compared to StyleGAN2 datasets. This observation is consistent with previous studies on landmark- and StyleGAN-based morph generation [22].

B. Morph Detection

Differential morph detector: We use FaceNet [16] as a verifier for our morphed images as shown in the Table I. We consider a positive pair a genuine image of a subject paired with a secondary bona fide instance of the subject, while a negative pair is a genuine image paired with a subject's respective morph. Verification results with a lower Area Under the Curve (AUC) and higher APCER values indicate that the morphs are successfully fooling the verifier. The morphed childrens' faces are able to fool FaceNet, with Equal Error Rate (EER) values over 9%. Across the three methods of morphing, StyleGAN2 consistently has a higher AUC then the landmark-based morphs. For example, while the Clarkson StyleGAN2 dataset has an AUC of 89.75%, the OpenCV and Facemorpher versions of the dataset have AUC 83.58% and 83.86%, respectively. This trend implies that FaceNet is able to differentiate between the morph and bona fide StyleGAN2 images at a higher rate than the landmark morph datasets. These results reinforce the known issue that StyleGAN2-generated morphs struggle to retain identity information at the same rate as the landmark-based morphs [3].

The verification results for the landmark morph dataset are significantly lower than FaceNet's expected morph detection performance. In [23], adult datasets are verified over 99% AUC using FaceNet. The morphed child datasets results in a significant AUC drop of approximately 16% when compared to adults. Additionally, there is a significant difference in performance of the verifier when comparing the older children in the UNCW and the young children found in the Clarkson dataset especially using the OpenCV method where the Clarkson OpenCV dataset has an AUC of 83.58% and the UNCW OpenCV dataset with an AUC of 92.23%.

Single morph detector: Using FaceNet [16], we train a binary classifier with a two-node output to detect morphs. The morph detector is trained on approximately 12,000 Facemorpher, OpenCV, and StyleGAN images of adult datasets. The detector learns the common artifacts of images using these morphing techniques. Table I shows the performance of the

TABLE II
MMPMR (%) FOR OUR SIX JUVENILE DATASETS.

Method		Facemorpher	OpenCV	StyleGAN2
Clarkson	FaceNet	91.31	87.98	73.82
	ArcFace	90.02	83.80	62.45
UNCW	FaceNet	99.32	97.87	90.40
	ArcFace	97.25	93.13	81.49

classifier on our six juvenile datasets. Similar to the differential scenarios, StyleGAN2 is shown to have a higher AUC in classification than the other datasets, specifically having an AUC of 79.68% for the Clarkson StyleGAN2 dataset and 92.66% AUC for the UNCW StyleGAN2 dataset, while their respective landmark morphs trend approximately 10% lower. The Clarkson landmark morphs and the UNCW landmark morphs all have APCER at BPCER=1% values above 93%, meaning that the morphs are effective at fooling the morph detector. In this scenario, we again observe the effects of aging in the performance of the classifier. The Clarkson dataset has a higher EER and lower AUC when across the methodologies. For the OpenCV morphs, Clarkson has an EER of 36.12% while UNCW has an EER of 24.77%. For Facemorpher, the EER for Clarkson is 33.54% and UNCW has an EER of 27.29%. This trend continues with StyleGAN2 having an EER of 27.06% and 14.55% for Clarkson and UNCW, which illustrates a bias toward the older children.

IV. Conclusion

In this paper, we generated high-quality morphed images from juvenile subjects. The morphed images were shown to retain their identity while being convincing enough to fool both single and differential morph detectors. While all datasets are shown to be effective at fooling morph detectors, the landmark-based morph images were more effective compared to StyleGAN2 morphs, which is consistent with adults datasets generated with the same methodology [3]. Across all morph detectors, morphed children pose a more significant threat than adult morphed datasets because of inherent bias when training deep learning models. This illustrated the necessity of further work to bridge the gap between facial recognition in adults and children as juvenile morphed images remain a threat to national security and child safety.

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] 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.

[2] Matteo Ferrara, Annalisa Franco, and Davide Maltoni, "The magic passport," in *IEEE International Joint Conference on Biometrics*, 2014, pp. 1–7.

[3] Eklavya Sarkar, Pavel Korshunov, Laurent Colbois, and Sébastien Marcel, "Vulnerability analysis of face morphing attacks from landmarks and generative adversarial networks," *CoRR*, vol. abs/2012.05344, 2020.

[4] Dana Michalski, Sau Yee Yiu, and Chris Malec, "The impact of age and threshold variation on facial recognition algorithm performance using images of children," in *International Conference on Biometrics*, 2018, pp. 217–224.

[5] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *IEEE international conference on automatic face & gesture recognition*, 2018, pp. 67–74.

[6] Nisha Srinivas, Karl Ricanek, Dana Michalski, David S. Bolme, and Michael King, "Face recognition algorithm bias: Performance differences on images of children and adults," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 2269–2277.

[7] Patrick Grother, Mei Ngan, and Kayee Hanaoka, "Ongoing face recognition vendor test (FRVT) Part 1: Verification," *National Institute of Standards and Technology*, 2018.

[8] Dana Kuefner, Viola Macchi Cassia, Marta Picozzi, and Emanuela Bricolo, "Do all kids look alike? evidence for an other-age effect in adults.," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 34, no. 4, pp. 811, 2008.

[9] John Davis, "Border crisis: CBP fights child exploitation," 2020.

[10] Alyssa Quek, "Facemorpher," Jan 2019.

[11] Satya Mallick, "Face morph using OpenCV - C++/Python," March 2016.

[12] Tero Karras, Samuli Laine, Miika Aittala, et al., "Analyzing and improving the image quality of StyleGAN," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2020, pp. 8110–8119.

[13] Naser Damer, Alexandra Moseguí Saladié, Andreas Braun, and Arjan Kuijper, "MorGAN: Recognition vulnerability and attack detectability of face morphing attacks created by generative adversarial network," in *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems*, 2018, pp. 1–10.

[14] Priyanka Das, Laura Holsopple, Dan Rissacher, Michael Schuckers, and Stephanie Schuckers, "Iris recognition performance in children: A longitudinal study," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 1, pp. 138–151, 2021.

[15] Karl Ricanek and Tamirat Tesafaye, "MORPH: A longitudinal image database of normal adult age-progression," in *IEEE international conference on automatic face and gesture recognition*, 2006, pp. 341–345.

[16] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

[17] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[18] International Organization for Standardization, "ISO/IEC DIS 30107-3:2016: Information technology biometric presentation attack detection," 2017.

[19] Ulrich Scherhag, Andreas Nautsch, Christian Rathgeb, et al., "Biometric systems under morphing attacks: Assessment of morphing techniques and vulnerability reporting," in *International Conference of the Biometrics Special Interest Group*, 2017, pp. 1–7.

[20] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

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

[22] Haoyu Zhang, Sushma Venkatesh, Raghavendra Ramachandra, et al., "MIPGAN-Generating strong and high quality morphing attacks using identity prior driven GAN," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 3, pp. 365–383, 2021.

[23] Kelsey O'Haire, Sobhan Soleymani, Baaria Chaudhary, Poorya Aghdaie, Jeremy Dawson, and Nasser M. Nasrabadi, "Adversarially perturbed wavelet-based morphed face generation," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2021, pp. 1–5.