Face Morphing Detection using Generative Adversarial Networks

Xinrui Yu, Guojun Yang and Jafar Saniie

Embedded Computing and Signal Processing (ECASP) Research Laboratory (http://ecasp.ece.iit.edu/)
Department of Electrical and Computer Engineering
Illinois Institute of Technology, Chicago, Illinois, USA

***Abstract*— Facial recognition system is used by various entities, including social media, identity verification, and security services. However, face morphing techniques present an adversarial challenge to the current facial recognition system. This is particularly the case for face morphing using Generative Adversarial Networks (GAN) method. To combat the face morphing we studied the GAN to detect morphed facial images. This has been achieved by using GAN to generate a large number of morphed images. Then, the morphed images are used to retrain the GAN for detecting the morphed images. The performance of this method is evaluated using facial images dataset and network structures.**

Keywords—Face morphing, face morphing detection, generative adversarial networks

I. INTRODUCTION

Facial recognition system plays an important role in modern society. Social media, identity verification, and security services rely on facial recognition systems to perform different tasks. While facial recognition algorithms based on biometrics can accurately recognize people in normal face images, they cannot distinguish morphed images from the real ones. This has created a major hazard for current facial recognition systems.

Among various methods to create morphed facial images, the Generative Adversarial Networks (GAN) is known to be very effective to the extent that the morphed images are real enough to confuse human [1]. In general, GAN consists of two competing neural networks, one network generates results similar to the sample images in the database, while the other network discriminates the morphed images from the sample images within the same database. After some iterations in this competition, the GAN method performs better compared to the design of a single neural network without competition. Therefore, this competing neural networks, GAN, can be trained to detect morphed facial images with acceptable levels of accuracy. As the face morphing neural network gets more and more powerful during the iterations, the discriminating neural network also gets increasingly more sensitive in detecting morphed images.

II. RELATED WORK

A. Face Morphing using GAN

Although GAN is a relatively new concept developed in 2014 [1], its usage rapidly expanded during the past few years, including implementations related to face morphing, image segmentation, and other image processing applications [2]. The results of face morphing using Deep-CNN [3] and GAN are real and unbelievable, generating discussion among researchers and ordinary people alike. An example of face morphing using GAN [4] is shown in Figure 1. The morphed faces show more aging.

Figure 1. Example of Face Morphing using GAN

Face morphing using GAN is not restricted to morphing from one person's face to another. Using someone's face image as a basis, it can also generate faces with additional/alternative features [4], such as adding/removing mustaches, aging, changing hair color, switching gender, and making the face more/less attractive. One successful method similar to GAN for face morphing is known as StarGAN. StarGAN is capable of performing image-to-image translations on multiple domains using only a single model, which leads to far more efficient training of multiple datasets. Furthermore, after training, it is possible to generate images of different target domains based on a single image, making face morphing faster and more efficient [4].

B. Detecting Face Morphing

The fast development and widespread of different face morphing methods present a serious challenge to current facial

978-1-7281-0927-5/19/$31.00 ©2019 IEEE

recognition systems based on biometrics. Because of the wide deployment of such systems in security applications such as border control, failing to recognize a morphed face image may lead to potential security threats.

In contrast to the potentially dire consequences, so far the challenge of developing a reliable morph detection method remains unsolved [5]. During the past few years, researchers have developed several methods to detect the morphed images. These methods are based on pattern recognition [5], Fourier spectrum analysis [6], and design of Deep-CNN (Deep-Convolutional Neural Network) [3]. While these methods showed desirable outcome, there is still room for further improvement in terms of accuracy and reliability.

III. Proposed Method for Detecting Morphed Images

A. Principles of Operation for GAN

The model of the neural networks used in this paper can be described as a multilayer perceptron. While this is true for most image processing related neural network, the most important difference between GAN and other neural network is the adversarial nature of GAN.

The GAN consists of two neural networks: one generative, the other discriminative. The generative network is similar to other neural networks used in image processing applications. It takes training images and target domain(s) as inputs, and output generated fake images. Conventional neural networks for face morphing end here. However, an adversarial network is needed for GAN. This is the discriminative neural network, in contrast to the generative neural network mentioned above. This network takes both the real and fake images as input and generates a predicted label. The label goes back to the generative neural network and is used in the next iteration to enhance its model. A diagram of an image processing GAN is shown in Figure 2.

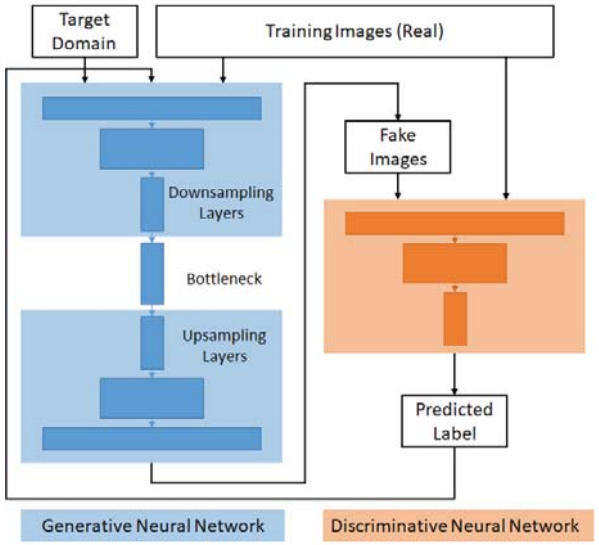

Figure 2. Structure Diagram of Image Processing GAN

The model of the neural networks used in this paper is called StarGAN and credited to [4]. For each training, multiple target domains can be given to the generative neural network, which made the training procedure more efficient.

The StarGAN used here can accept multiple domains as input. However, the labels for these domains must be predefined in the dataset data structure. The dataset used in this paper is CelebFaces Attributes Dataset (CelebA), which is a large facial image and attribute dataset with 202,599 facial images of 10,177 people [7]. The resolution of these images (178×218) is enough for various applications. What is particularly important and essential is that each image in this dataset has 40 binary attribute annotations. The attributes include hair colors, items wore, facial hair, etc. For example, one with brown hair will have the annotation of 1 for brown hair attribute, and -1 for black hair and blonde hair attributes. These annotated attributes gave us the potential to use them to generate morphed images based on a certain attribute, like changing one's hair color, making someone more attractive and so on.

The StarGAN generates morphed images during the training process. As the number of iterations goes up, the morphed images get more difficult to distinguish from the real ones. This is shown in Fig 3. Five face attributes are chosen in this example, they are attractive, blonde, bald, gender, and age. The result corresponds to the negative attribute annotation; if the attribute is contradictory with other attributes, like brown hair, the result will be the closest match of the corresponding attribute. For example, the five attributes' annotations of the real image are respectively -1, 1, -1, 1, -1, which means that the one in the image is not attractive, has blonde hair, not bald, male, and old. The result will be attractive, blonde, bald, female, young for each domain respectively, as shown in the columns of Figure 3. The rows are results from different numbers of iterations. As shown in Figure 3., the fake images are becoming more and more indistinguishable as the number of iterations goes up. In general, the model after training on the CelebA dataset is capable of generating morphed images which are difficult to detect with even trained human eyes.

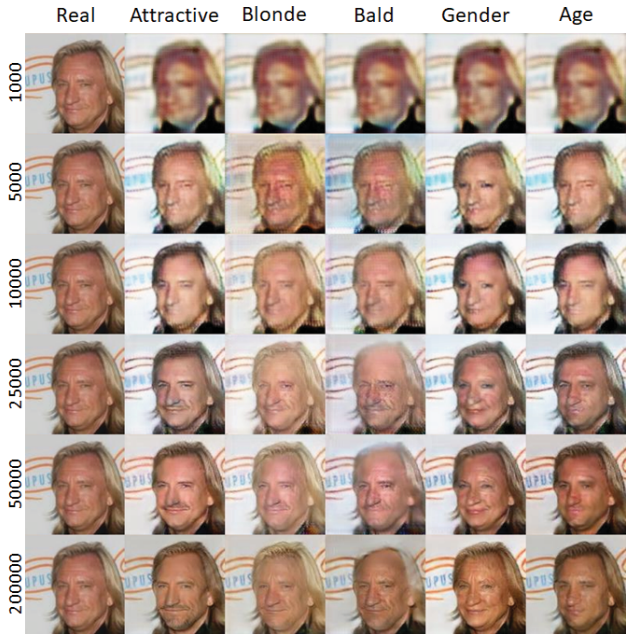

Figure 3. Image Morphing Result of Different Attributes and Iterations

B. Morphed Image Detection Using StarGAN

As the leading trend in the field, face morphing using GAN is becoming more and more widespread than ever before. These morphed images generated using GAN can be unrecognizable even for human workers/researchers in this field and are also challenging for conventional methods of detection mentioned in the previous section. However, because of the adversarial nature of GAN, for every successful generative neural network, there is also a corresponding discriminative network. As the images generated by the generative neural network get harder to be distinguished, the discriminative capability of the discriminative network also gets better. Theoretically, even if the morphed images cannot be detected reliably by any other methods, the discriminative neural network always retains a certain level of confidence detecting these morphed images.

To use StarGAN to detect morphed images, the dataset structure is modified. The generated images together with part of the original training images are given an extra attribute, real. The annotation of this attribute is -1 for all morphed images and 1 for all original training images. This new image set is used as the training set for morphed image detection. The model of the retrained StarGAN is capable of detecting morphed images.

After training, some untrained morphed and original images will be given to the model as testing images. They will all be treated as if they are morphed images ("real" attribute = -1) and be used to generate pseudo originals of the morphed image. Then the error compared with real original is calculated by the discriminative part of StarGAN. If this pseudo original has a large error compared to the real original, this shows that the generated originals differ greatly from the testing image and the image will be determined to be a morphed image.

Generally, neural networks are trained on a specific type of domain/feature-level. For example, a neural network specialized in creating aged face image will only be trained according to a single age tag assigned to the face images in the training set, and will only be able to generate and distinguish images with varying ages. This is rather inconvenient for our task, as the morphed images can be of many different feature levels. To detect such image with conventional neural network, one needs to train multiple neural networks of different domains, and the time and effort spent on training and modifying so many neural networks will make this method unworthy. The StarGAN provided a solution to such dilemma. It is capable of performing image-to-image translations on multiple domains using only a single model, therefore greatly reducing the time and effort to be spent in training and modifying the neural network. By training on multiple domains simultaneously, it is possible to use a single model after extensive training to deal with multiple types of morphed images.

IV. Implementation and Results

A. Generate Morphed Images

For experimental studies two face images of our research laboratory members are used to generate morphed images. The domains used in this training are respectively: Attractive, Chubby, Bald, Gender, and Age. The images are generated using the GAN after 200,000 iterations. The software environment used is Python 3.7, PyTorch v1.0.1, and TensorFlow 1.12. The total training time for 200,000 iterations is about 21 hours on a single NVIDIA GTX1070 GPU. The GPU is CUDA compatible. Increasing the resolution of the images from 128×128 pixels to 176×176 pixels increases the training time to about 31 hours. The original and morphed images are shown in Figure 4. Qualitative evaluation is performed on these morphed images. While the bald images look suspicious with some color overflow, the gender and age change images look quite credible, especially to someone who has never seen these faces in real life.

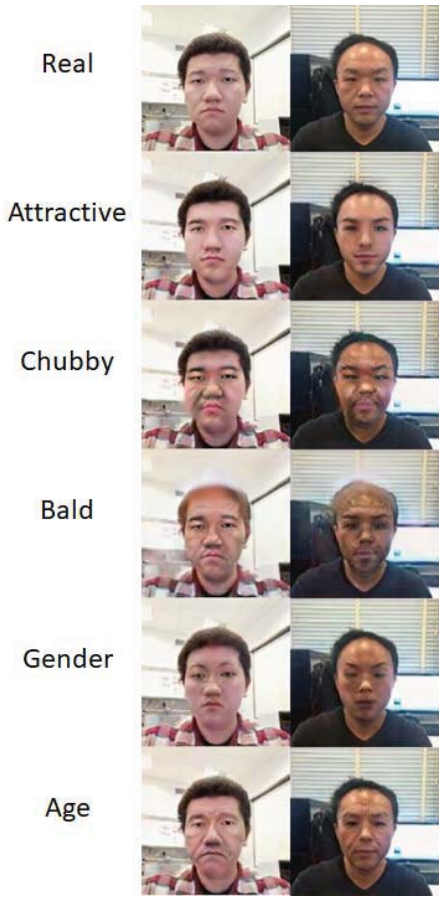

Figure 4. Image Morphing Results of Two Lab Members

A quantitative evaluation is performed by calculating reconstruction error compared with the real/original image. The reconstruction error of the model on different numbers of iterations is shown in Figure 5. The calculation of reconstruction error is [4]:

$$E_{Rec} = E_{x,c,c_0}[\|x - G(G(x,c),c_0)\|] \quad (1)$$

Where x is the original image, G is the operation of the generative neural network, $G(x,c)$ is the morphed image obtained from original image x and desired domain label c. Similarly, $G(G(x,c),c_0)$ is the reconstructed original image obtained from the morphed image $G(x,c)$ and original domain label c_0. L1 norm is used to calculate the error in this formula,

and the result can be seen as an error percentage compared with the original image.

From Figure 5, we can see that the error decreases as the number of iterations increases. The x-axis shows the number of iterations, and the y-axis shows the reconstruction error in percentile. The error average over 100 and 500 iterations are also shown in this figure. The rate at which the error decreases with the number of iterations is consistently dropping. It is worth mentioning that the error is not monotonically decreasing even when using an average over 500 iterations. This is expected because of the complex nature of the neural network. Also, Figure 5. shows that the reconstruction error converges to less than 5% above 15,000 iterations.

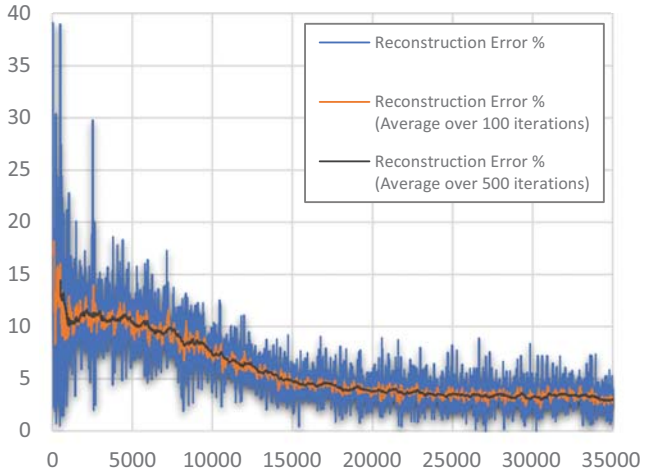

Figure 5. Reconstruction Error of Different Number of Iterations

B. *Morphed Image Detection*

To detect morphed face images, the GAN needs to be trained with a dataset containing a respectable number of morphed face images and their corresponding attributes. This is done by adding some of the generated images in the previous section to the dataset, replacing some of the original images. The morphed images are reshaped and resized to match the format of the original images in the dataset. A new attribute is added to the attribute annotation table, "real". A value of -1 is assigned to the morphed images for this attribute, and the real images are assigned a value of 1 for this attribute.

For testing, 100 morphed images generated by the StarGAN are used. They are generated from 10 facial images, each morphed with 10 different attributes. Among the 100 morphed images, 72 are determined to be morphed by the network. As these images are generated by GAN and considered hard to detect even with human eye [4], the detection of morphed images with 72% accuracy is promising. The result of morphed images from other sources is somewhat less conclusive. True positive of different groups varies greatly. For example, over 95% of images morphed using distorting mirror effect is detected correctly.

V. Conclusion

According to the test results mentioned above, we can detect a morphed face image in real-time with reasonable accuracy. The research in this paper would be beneficial for the development of a reliable and accurate method to distinguish morphed face images. By providing more morphed face images of different origins like [8], it is possible to make this method more robust and reliable. A GAN for a specific facial attribute, such as facial aging [9], might be used as a basis to further improve the detection capability. Furthermore, a dataset containing face images of different viewing angles like RaFD [10] can further increase the effectiveness of the method, making it possible to identify morphed face images of a different viewing angle. The result and experience gained from this research can be researched and applied to video processing for security applications.

References

[1] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. "Generative Adversarial Nets," *Advances in Neural Information Processing Systems*, 2014, pp. 2672-2680.

[2] X. Zhang, X. Zhu, 3. X. Zhang, N. Zhang, P. Li and L. Wang, "SegGAN: Semantic Segmentation with Generative Adversarial Network," *2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM)*, Xi'an, 2018, pp. 1-5.

[3] R. Raghavendra, K. B. Raja, S. Venkatesh and C. Busch, "Transferable Deep-CNN Features for Detecting Digital and Print-Scanned Morphed Face Images," *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Honolulu, HI, 2017, pp. 1822-1830.

[4] Y. Choi, M. Choi, M. Kim, J. Ha, S. Kim and J. Choo, "StarGAN: Unified Generative Adversarial Networks for Multi-domain Image-to-Image Translation," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, 2018, pp. 8789-8797.

[5] U. Scherhag, C. Rathgeb and C. Busch, "Towards Detection of Morphed Face Images in Electronic Travel Documents," *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, Vienna, 2018, pp. 187-192.

[6] L. Zhang, F. Peng and M. Long, "Face Morphing Detection Using Fourier Spectrum of Sensor Pattern Noise," *2018 IEEE International Conference on Multimedia and Expo (ICME)*, San Diego, CA, 2018, pp. 1-6.

[7] Z. Liu, P. Luo, X. Wang and X. Tang, "Deep Learning Face Attributes in the Wild," *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, 2015, pp. 3730-3738.

[8] T. Karras, S. Laine and T. Aila, "A style-based generator architecture for generative adversarial networks." *arXiv preprint arXiv*:1812.04948, 2018

[9] X. Wang, Y. Zhou, D. Kong, J. Currey, D. Li and J. Zhou, "Unleash the Black Magic in Age: A Multi-Task Deep Neural Network Approach for Cross-Age Face Verification," *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, Washington, DC, 2017, pp. 596-603.

[10] O. Langner, R. Dotsch, G. Bijlstra, D. H. J. Wigboldus, S. T. Hawk, and A. V. Knippenberg, "Presentation and validation of the Radboud Faces Database," *Cognition & Emotion*, vol. 24, no. 8, 2010, pp. 1377–1388.