FRONTAL FACE GENERATION FROM MULTIPLE POSE-VARIANT FACES WITH CGAN IN REAL-WORLD SURVEILLANCE SCENE

Zhu-Liang Chen Qian-Hua He Wen-Feng Pang Yan-Xiong Li*

School of Electronic and Information Engineering
South China University of Technology, Guangzhou, China

ABSTRACT

It is well known that frontal face is much easier to be recognized than pose-variant face for both human and machine perception. However, it is not easy to acquire a frontal face in real-world video surveillance.

This paper proposes a method to synthetize a frontal face for recognition in video surveillance scene, which is based on Conditional Generative Adversarial Networks (cGAN) with input of multiple pose-variant faces from a video. Experimental results show that the proposed approach can generate suitable frontal faces and improve face recognition by around 20% on a dataset of 43276 face images from 19 persons, collected from the real-world video surveillance scene. The effectiveness of multiple frames against single frame as input is demonstrated. Moreover, we investigate the generator with different depth for synthetizing frontal faces, in which an up-down sampling trick is designed for synthetizing higher quality frontal face images and boosts the performance of the generator.

Index Terms—synthetize face, multiple frames input, cGAN, video surveillance

1. INTRODUCTION

The face recognition problem always attracts much attention from academia and industries because of its wide applications including surveillance, identity authentication and so on [1]. From hand-crafted features to deep learning methods, the performance of face recognition has been greatly improved and exceeded the human's level in some face recognition databases [2][3][4]. Nevertheless, it still has a large gap to reach human's level performance in the real-world scene as a result of low quality face images (blur, low resolution, variant poses).

First of all, we define those video clips containing a person passing the surveillance camera as motion events. Face recognition is implemented in every motion event clip instead of every single face image frame from video.

It is unreliable that only one ill-posed face is used to synthetize the frontal face independently for a face recognition task. Fortunately, the multiple relevant face images in one motion event have variant poses containing same environmental information and sufficient facial features. Hence, they are reasonably used to synthetize one reliable frontal face for improving recognition performance.

This paper will focus on the face recognition problem in real-world surveillance video scene with low-quality face images rather than public high-quality face databases. An approach is presented to synthetize frontal face for improving face recognition based on cGAN [5][6]. In details, the cGAN can synthetize aligned frontal face using 3 frames face images from one motion event with the same identity as input even though the face images have variant poses.

The cGAN has a generator and discriminator. The generator consists of encoder-decoder architecture [7], and can synthetize frontal face images to fool the discriminator. The discriminator based on the deep convolutional generative adversarial networks (DCGAN) [8], is trained to evaluate whether they are real or fake face images. This discriminator focuses on the local patches for classification instead of the global [9]. The condition of this cGAN is L1 loss between synthetic output image and target image, which can send the feedback to generator and make the synthetic face images more realistic.

Due to input images with low resolution (40x40 pixels), the cGAN will cause blurry face images if we design the same size (40x40 pixels) of synthetic images at generator. In order to deal with this problem, a suitable generator of cGAN and one up-down sampling trick are presented to generate higher quality face images. Following this up-down sampling trick, input images and target images are implemented up sample to 256x256 pixels and sent to the cGAN. Next step, the cGAN synthetizes 256x256 pixels frontal face which is implemented down sample into 40x40 pixels later for face recognition. This solution makes the generator synthetize higher dimensions data containing more features and boost the performance of generator.

2. PREVIOUS WORK

* eeqhhe@scut.edu.cn;
The work was supported by the National Nature Science Foundation of China (Grant No. 61571192, 61771200);

In recent years, the performance of face recognition has been significantly improved as a result of deep learning. However, in practice, we always capture pose-variant faces with low resolution in video surveillance, which is still a great challenge for recognition. In the pose variation challenge, one solution is to adopt hand-crafted features [10][11], while another one tries to synthetize, rotate or align to obtain a frontal face from the original image [12][13][14].

The hand-crafted methods always play well in one specific case instead of general scenes, which have weak generalization ability. Recently, GAN methods are used to synthetize face images from random noise [15]. Particularly, TP-GAN is presented to utilize one single variant pose face image to synthetize frontal view face [16]. But the recognition performance of the variant pose face with large deflection angle (more than a 75°) cannot satisfy practical applications. Most of these methods based on the face image datasets with high resolution (more than 100x100 pixels) instead of real-world scenes datasets, which are inconsistent with practical situation either.

3. APPROACH

3.1. Objective

In real-world video surveillance scenes, face recognition is always implemented in one motion event as a basic unit. Obviously, a series of face images with variant poses can be easily captured, and these pose-variant face images have effective and sufficient information to synthetize a semantically frontal face. This can be considered as fusing features from variant pose face images. Hence, the cGAN in this paper is adopted to synthetize frontal face with multiple variant pose face images from one motion event.

Generally, GAN model cannot generate precise and reliable output only depending on input images. In contrast, GANs with some condition settings could generate a reasonable result [17]. In this paper, the cGAN uses L1 loss between target images and synthetic output face images as the condition setting. The objective functions of the cGAN are as follows:

$$L_G = E\left[\log(1 - D(x, G(x,z)))\right] + \lambda E\left[\left\|y - G(x,z)\right\|_1\right] \quad (1)$$

$$L_D(D,G) = E[\log(D(x,y))] + E[\log(1 - D(x, G(x,z)))] \quad (2)$$

$$L_{cGAN} = \min_G \max_D \left[L_{cGAN}(D,G) + \lambda E\left[\left\|y - G(x,z)\right\|_1\right]\right] \quad (3)$$

The cGAN is trained on adversarial strategy. Restricted by these objectives, the discriminator ***D*** wants to maximize this objective while the generator ***G*** tries to minimize it. Due to the L1 loss between target face image and synthetic output image, this objective will make the output image close to the target image depending on weight lamda. The ***G*** is trained by formula (1) while the ***D*** is train by formula (2).

L1 loss is adopted since it encourages less blurring than L2 [18]. Under the L1 loss constraint condition, the cGAN will be guided to generate an aligned frontal face with the same identity because the target face image is an aligned frontal face and the input images are from same identity as well.

3.2. Up-down sampling trick for boosting cGAN and recognition performance

Some past research works showed that low-quality face images would substantially drop down the performance of face recognition. As a result, the majority of face datasets have at least 100x100 pixels face images to guarantee clear faces with more details and features. However, in our dataset, whose face images extracted from real-world video surveillance always only have 40x40 to 60x60 pixels. Therefore, it doesn't make sense to use high resolution (more than 100x100 pixels) face images datasets to build our practical model and apply it in real-world scenes.

Thus, an up-down sampling trick is presented to boost the performances of both the generator and the recognition. Concretely, the input and target images are implemented up-sampling (bilinear interpolation) operation from 40x40 to 256x256 pixels. After this pre-processing, the cGAN will synthetize 256x256 pixels frontal face images. In the next step, the synthetic face images will be down-sampled to 40x40 pixels image for face recognition.

This solution looks ordinary but very powerful, which will be validated in experiment section. The generator has stronger potential to learn higher dimensional features and synthetize better face images, which is coincident with the original goal.

3.3. Conditional Generative Adversarial Network

3.3.1. Architecture of Generator

The generator ***G*** consists of encoder and decoder, like the architecture of Auto-Encoder. The input of generator is three random frames containing gray scale face with variant pose in motion event and the generator can synthetize one RGB frontal face images.

The generator with deeper layers can obtain stronger learning capability. Moreover, the deeper generator will generate face images in a larger size. Through changing the depth of generator, the generator can regulate the size of output face images. Therefore, the input images will be implemented up sampling to 256x256 pixels, and the generator will use deeper layers to boost its performance for synthesizing 256x256 pixels face images, which is the up sample part of the solution presented at section 3.2. The architecture of generator is shown at Figure 1.

The encoder of generator has 8 basic modules and the number of filters of each module is 64-128-256-512-512-512-512-512. The decoder is at reverse order. Each module of encoder is Conv-BatchNorm-leakyRelu form, while the decoder has Deconv-BatchNorm-Relu form.

Those modules can speed up training [19]. Notice, one

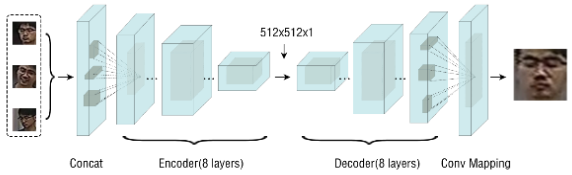

Figure 1. The architecture of generator.

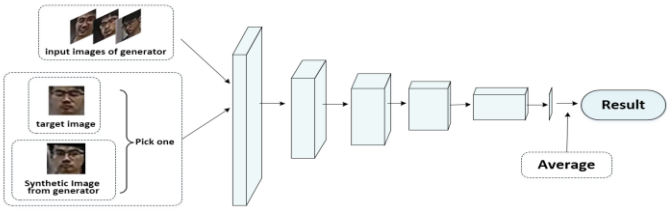

Figure 2. The architecture of discriminator.

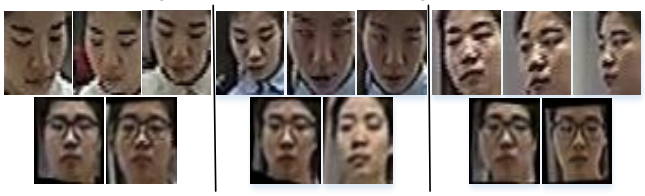

Figure 3. Every column is one sample. The images at the first row are input variant-pose faces. The left one at the second row of every column is synthetic face and the right one is target image

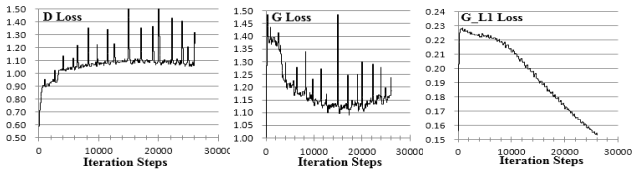

Figure 4. The first image is loss of discriminator, the second one is loss of generator and the last one is loss of L1.

256x256x3 convolution layer will be both added at beginning of encoder for input image and at the end of decoder followed by Tanh function for generating output face images with 3 RGB channels as well.

3.3.2. Architecture of Discriminator

Discriminator has 5 basic modules and the number of filters of each module is 64-128-256-512-1. The basic module of discriminator is Conv-BatchNorm-leakyRelu form. The architecture of discriminator is shown at Figure 2. The input of discriminator has two kinds of image pairs. One pair is 3 input images of generator and the target frontal face image with the same identity. Another pair is 3 input images of generator and its synthetic face image. These two pairs will be individually sent to the discriminator as input and get two scores. The sum of these two scores is regarded as the loss of discriminator, which is shown at formula (2). The input of ***D*** is pair-image instead of single target or synthetic image because these two pairs both have the input of ***G***, which can bring more noise to avoid overfitting and let the ***G*** and ***D*** acquire more poses-variant information from input images.

More specifically, the input pairs will be both implemented up-sampling trick to 256x256 pixels. At the last layer of discriminator, the discriminator will output the 30x30x1 dimensions data where each element (its value is between 0 and 1) is represented the confidence of its receptive field. The receptive field focuses on local features and the average value of these 30x30x1 data is regarded as the score mentioned above. Such an effective patch discriminator was previously explored in [9][18][20][21].

4. EXPERIMENTS

4.1. Motion event video clips dataset for face recognition from real-world surveillance scene

In video surveillance, it is difficult to capture high resolution face images. Those models trained on high resolution face image database cannot guarantee the generalization ability in real-world scenes. we collected one person entrance video dataset containing 3465 clips, which has 43276 face images extracted from 19 identities. Each video clip is one person entrance event lasting 8 to 10 seconds. The resolution of face images in these clips is 40x40 to 60x60 pixels. All experiments in this paper are based on this dataset.

4.2. Training cGAN

The input of ***G*** is 3 face images from one single person entrance event video clip. The one kind of input pairs of ***D*** is the 3 input images of ***G*** and one randomly selected target image, the other is the 3 input images of ***G*** and its synthetic face image. The target images are frontal face images, which are selected manually and aligned by affine transformation using 68 points facial landmarks [22].

The randomicity of combining target image and input image can be considered as a kind of data augmentation and regularization methods. In addition, it adds some random noise into the model at the training phase through this method. Hence, it can reduce the variance of the cGAN to improve the generalization performance.

The cGAN is trained with 50 epochs and 60 batch sizes based on our dataset. The weight lamda of L1 loss is 100 at the first 30 epochs and 150 at the last 20 epochs. The scale of training set and test set from our dataset is 6:4. At face synthesis part (test part), it takes 55.876 ms approximately to synthetize one frontal face image in the computer with Intel Core I7 6700 (3.4GHz) and GeForce GTX TITAN X.

4.3. Face synthesis and face recognition

From Figure 3, we can see three cases of the input images, synthetic output images, and target images. Besides, the loss of generator, discriminator and L1 are shown at Figure 4, which represent the learning process.

Table 1 Face recognition of different images under different settings, the resized setting means it will be resized into 40x40 pixels.

Image	LBP				Resnet-DLIB			
	KNN-1	KNN-3	KNN-5	SVM	KNN-1	KNN-3	KNN-5	SVM
Input Images	24.6%	26.9%	29.4%	11.2%	63.8%	67.1%	70.9%	80.0%
Synthesis from single-frame cGAN	33.8%	36.2%	38.8%	11.3%	75.2%	78.1%	79.9%	82.2%
Synthesis from Ours, NO Resized	27.4%	29.1%	35.4%	18.1%	85.8%	88.2%	88.7%	90.1%
Synthesis from Ours, Resized	33.6%	36.0%	40.1%	17.8%	89.5%	89.9%	90.7%	93.2%
Target Image	56.7%	56.5%	56.5%	10.9%	99.1%	99.0%	99.4%	98.8%

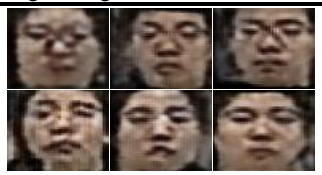

Figure 5. Six synthetic frontal faces from weakened generator under the same setting of our approach. The synthetic images are blurry and it is hard to distinguish who she is, even though they have facial features. Actually, the identity of these images is the same as Figure 3.

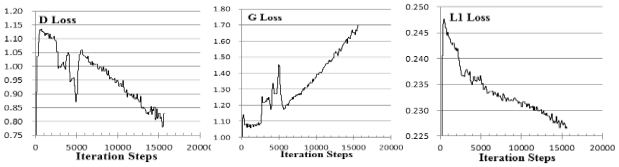

Figure 6. The loss of discriminator, generator and L1 of the cGAN with weakened generator under the same setting of our approach. These losses represent that this model cannot generate good face images to fool the discriminator.

KNN (with different k-neighbors) and SVM (kernel= liner, C=1) are used for classification tasks as the benchmark. KNN with 1-neighbor is regarded as maximum Euclidean distance method. Uniform LBP and Resnet-Dlib-model [23] are respectively used as features for face recognition. In this part, the up-down sampling trick is applied at the cGAN. The input, output and target face images are both validated under the same settings of classification algorithms. The recognition results are shown at Table 1.

First of all, the traditional uniform LBP feature get unacceptable recognition performance, which demonstrates the difficulty of this surveillance video scene. Thus, Resnet-DLIB features is adopted to validate effect of the cGAN.

The original selective frontal faces obtain the best recognition performance without any doubt. Comparing synthetic frontal faces with pose-variant faces, the synthetic frontal faces are better than pose-variant faces, regardless of KNNs or SVM. Impressively, the cGAN can improve recognition performance by about 20% in some cases.

To demonstrate the different effect of multiple frames and single frame at the input layer, one single-frame cGAN is designed under the same architecture as the cGAN, but the input of this model is a single face image with variant pose. This single-frame cGAN is trained under the same setting as our cGAN. As the results shown at Table 1, the multiple-frame cGAN obviously acquires more features and obtains nearly 10% improvement of recognition comparing with the single-frame cGAN. As a comparison, TP-GAN using single face image to synthetize frontal face at Multi-PIE dataset with setting two got 98.68%, 98.06%, 95.38%,87.72%, 77.43% and 64.64% rank-1 recognition rate under the 15°, 30°, 45°, 60°, 75° and 90° cross views [16].

4.4. Up-down sampling solution and the validation of generator with different depth for generating different resolution face images

There are 2 methods to generate 40x40 resolution face images. One method is the up-down sampling solution we presented above. Another method is to synthetize 40x40 face image directly by the generator with less layers. As the validation, we removed the last 2 layers at encoder and the first 2 layers at decoder to generate a 64x64 pixels image. The training results of this model (weakened generator) are shown at Figures 5 and 6. From the quality of the synthetic face images and its loss results, this weakened generator cannot successfully learn the features, since less layers represents a weaker learning capability. Hence, the up-down sampling solution and one generator with suitable depth are adopted to deal with this challenge.

Furthermore, due to the original input images with low resolution and limited information (40x40 pixels), it doesn't make sense to use 256x256 pixels synthetic image directly for recognition because it has extra noise with super resolution. Hence, down-sampling is implemented at the synthetic images to 40x40 pixels, which can reduce noise to improve 2 % recognition performance as the Table 1 shown. This denoising performance is another benefit of the up-down sampling solution.

5. CONCLUSION

We present an approach using cGAN to synthesize frontal face with 3 face images randomly selected from one motion video with low resolution (40x40 pixels). The synthetic frontal face images can improve the performance of face recognition by about 20% on the real-world video surveillance dataset. Moreover, the effectiveness of multiple frames in video analysis is demonstrated through comparing with the single frame input. One up-down sampling solution is presented to deal with low resolution face images challenge, which significantly boosts the performance of generator and face recognition.

6. REFERENCES

[1] Michel Owayjan, Amer Dergham, Gerges Haber, Nidal Fakih, Ahmad Hamoush, and Elie Abdo, "Face recognition security system", *New Trends in Networking, Computing, E-learning, Systems Sciences, and Engineering*, pp. 343–348. Springer, 2015.

[2] Sun Y, Wang X, Tang X, "Deep learning face representation from predicting 10,000 classes", *Computer Vision and Pattern Recognition (CVPR), IEEE Conference*, pp. 1891-1898, 2014.

[3] Schroff F, Kalenichenko D, Philbin J. "Facenet: A unified embedding for face recognition and clustering", *Computer Vision and Pattern Recognition (CVPR), IEEE Conference*, pp. 815 - 823, 2015.

[4] Sun Y, Wang X, Tang X. "Deeply learned face representations are sparse, selective, and robust", *Computer Vision and Pattern Recognition (CVPR), IEEE Conference*, pp. 2892 - 2900, 2015.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D.Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. "Generative adversarial nets", *Conference on Neural Information Processing Systems (NIPS)*, pp. 2672-2680, 2014.

[6] M. Mirza and S. Osindero. "Conditional generative adversarial nets", *arXiv:1411.1784*, 2014.

[7] G. E. Hinton and R. R. Salakhutdinov. "Reducing the dimensionality of data with neural networks", *Science,* pp. 504–507, 2006.

[8] A. Radford, L. Metz, and S. Chintala. "Unsupervised representation learning with deep convolutional generative adversarial networks", *arXiv:1511.06434*, 2015.

[9] C. Li and M. Wand. "Precomputed real-time texture synthesis with markovian generative adversarial networks", *European Conference on Computer Vision (ECCV)*, pp. 702-716, 2016.

[10] D. Chen, X. Cao, F.Wen, and J. Sun. "Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification", *Computer Vision and Pattern Recognition (CVPR), IEEE Conference*, pp. 3025-3032, 2013.

[11] D. J. Rezende, S. Mohamed, and D. Wierstra. "Stochastic backpropagation and approximate inference in deep generative models", *arXiv:1401.4082*, 2014.

[12] H. Zhang, Q. Li, and Z. Sun. "Combining data-driven and model-driven methods for robust facial landmark detection", *arXiv:1611.10152*, 2016.

[13] Z. Zheng, L. Zheng, and Y. Yang. "Unlabeled samples generated by gan improve the person re-identification baseline in Vitro", *arXiv:1701.07717*, 2017.

[14] Xavier Fontaine, Radhakrishna Achanta, and Sabine Susstrunk, "Face Recognition in Real-world Images", *International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE Conference*, 2017.

[15] David Berthelot, Thomas Schumm, Luke Metz, "BEGAN: Boundary Equilibrium Generative Adversarial Network", *arXiv:1703.10717*, 2017.

[16] Rui Huang, Shu Zhang, Tianyu Li and Ran He, "Beyond Face Rotation: Global and Local Perception GAN for Photorealistic and Identity Preserving Frontal View Synthesis", *arXiv:1704.04086,* 2017.

[17] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, Alexei A. Efros. "Context Encoders: Feature Learning by Inpainting", *Computer Vision and Pattern Recognition (CVPR), IEEE Conference*, pp. 2536-2544, 2016.

[18] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks", *arXiv:1611.07004*, 2016.

[19] Sergey Ioffe, and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift", *International Conference on Machine Learning (ICML)*, pp. 448-456, 2015.

[20] A. A. Efros and T. K. Leung. "Texture synthesis by nonparametric sampling", *International Conference on Computer Vision (ICCV), IEEE Conference*, pp. 1033–1038, 1999.

[21] L. A. Gatys, A. S. Ecker, and M. Bethge. "Texture synthesis and the controlled generation of natural stimuli using convolutional neural networks", *arXiv:1505.07376*, 2015.

[22] Vahid Kazemi and Josephine Sullivan, "One Millisecond Face Alignment with an Ensemble of Regression Trees", *Computer Vision and Pattern Recognition (CVPR), IEEE Conference*, pp. 1867-1874, 2014.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep Residual Learning for Image Recognition", *Computer Vision and Pattern Recognition (CVPR), IEEE Conference*, pp. 770-778, 2016.