A Static2Dynamic GAN Model for Generation of Dynamic Facial Expression Images

Yoshiharu Kawai, Masataka Seo, and *Yen-Wei Chen
Graduate School of Information Sci. and Eng., Ritsumeikan University, Siga, Japan
Corresponding authors: chen@is.ritsumei.ac.jp

Abstract—Facial expression generation, especially dynamic facial expression generation from a static natural (expressionless) face image, plays an important role in the fields of entertainment, game, and social communication. Several approaches based on machine learning including deep learning techniques have been developed or proposed for facial expression generation. However, most of them are focused on static facial expression generation. In this paper, we propose a static-to-dynamic model (static2dynamic) based on a 3D conditional generative adversarial network. Herein, the dynamic facial expression image is treated as a 3D image. The effectiveness of the proposed method is demonstrated by generating dynamic smile and angry facial expression images.

Keywords—Static2dynamic, Generative Adversarial Nets, U-net, dynamic facial expression, Pix2Pix, 3D

I. INTRODUCTION

Facial expression generation, especially dynamic facial expression generation from a static natural (expressionless) face image, plays an important role in the fields of entertainment, game, and social communication. Several approaches based on machine learning including deep learning techniques have been developed or proposed for facial expression generation [1-3]. Some of these methods are landmark-based facial deformation or morphing techniques. The generation performance is dependent on the accuracy of the landmark detection, thus it is difficult generating a real expression texture. We proposed a method without the use of facial landmarks based on a conditional generative adversarial network (CGAN) [3], which is also known as pix2pix, for facial expression generation in our previous study [4]. However, these methods are focused on static facial expression generation. There is little research on dynamic facial expression generation. Inspired by our previous work, we propose a static-to-dynamic model (static2dynamic) based on a 3D conditional generative adversarial network (3D CGAN) to generate dynamic facial expressions for only a static natural (expressionless) face image. In this proposed method, the dynamic facial expression image is treated as a 3D image. Also, this method is without the use of facial landmarks. We demonstrated the effectiveness of the proposed method by generating dynamic smile and angry facial expression images.

II. METHOD

In this section, we first describe the static facial expression generation method using pix2pix [3] as related work. Then, we describe our proposed static-to-dynamic (static2dynamic) model based on a 3D CGAN, which is an extension of the conventional 2D CGAN, for the dynamic facial expression generation.

A. Static facial expression generation based on pix2pix

Generative adversarial network (GAN) is a generative model that learns a mapping from a random noise vector z to an output image. CGAN learns a mapping from observed image x and random noise vector z to y, i.e., G: {x, z} \rightarrow y.

Pix2pix builds a generation model by learning two neural networks: generator and discriminator [3]. The generator is trained to produce output images which cannot be distinguished from a real image by an adversarial training discriminator. The discriminator is trained to do as well as possible at identifying whether the generator has a fake image or image from training ground truth data. However, the generator is trained until the discriminator cannot identify whether real or fake. This network learns this task in a supervised manner using CGAN. To generate real and beautiful images, pix2Pix learns the mapping from an input image to output image, as well as, a loss function to train this mapping. The generator is based on U-Net. The U-Net architecture allows low-level information such as edge information to short cut across the encoder and decoder, and for our discriminator, we use a convolutional patch GAN classifier which only penalizes structure at the scale of image patches.

The overall flow of the previous static expression generation model [4] is shown in Fig. 1. Its generator and discriminator architectures are shown in Fig.2(a) and 2(b), respectively.

![Overall flow of static facial expression generation based on Pix2Pix](image)

**Fig.1.** The overall flow of static facial expression generation based on Pix2Pix
Inspired by pix2pix, we propose a static2dynamic model based on a 3D CGAN to generate dynamic facial expressions for only a static natural (expressionless) face image. In this method, the dynamic facial expression image is treated as a 3D image. The conventional CGAN is extended to a 3D CGAN, which consists of 3D convolution layers, 3D pooling layers, and 3D deconvolution layers. The dynamic features (temporal features) are extracted by 3D convolution layers together with the 2D spatial features. The dynamic images are generated (reconstructed) by 3D deconvolution layers. Although we only use one static natural (expressionless) image (input image), we copy the input 2D static image to generate a time-series image with the same number of frames as output dynamic images, which is used as a 3D input to the static2dynamic model (network). The overflow of the proposed static2dynamic GAN model is shown in Fig. 3. The generator and discriminator architectures are shown in Fig. 4(a) and (b), respectively. We adopt the L1 loss function to the generator, which is shown in Eq.(1).

\[ L_{L1}(G) = E_{x,y \sim P_{data}}[||y - G(x)||] \] (1)
Fig 4 (a) Generator and (b) discriminator of the proposed static2dynamic GAN model.
III. EXPERIMENTS

In this section, we first describe the details and experimental results of static expression generation using pix2pix [4] as a related work. Then, we describe the details and experimental results using our proposed static-to-dynamic (static2dynamic) model.

A. Pix2Pix Dataset and Implementation

To train the generator model, we took a facial image in a dark room at our laboratory to create pairs of expressionless face image and expression face image. Our dataset consists of pairs of expression images and expressionless of 15 male and 15 female individuals. Each individual has two expressions: smile and angry. Each expression starts from the expressionless (natural) image to the expression image. Figs 5(a), 5(b), and 5(c) show examples of expressionless, angry, and smile, respectively. Fig 6 shows an image in which a 500 × 500 pixels ROI region is extracted from all the images with the center of the eyebrow as the center of the ROI. This preprocessing eliminates blurring of the generated image. We augmented data by using translations to increase the training dataset. After selecting an expressionless image and an expression image as a pair, we translate both images with a random value Δ (−20 < Δ < +20) in the x-direction and y-direction as shown in Fig.7. Note that the expression image and its corresponding expressionless image must be registered in the preprocessing. Otherwise, blurring may occur in the generated image. The images were also resized to the image size (256 × 256 pixels) we want to generate. All images generated by preprocessing are used as a pair. Finally, in our experiment, we augmented 336 pairs of images for each expression per person. Hence, we have 10080 pairs as training data for each expression network.

The testing data is an expressionless image of a female, which is not included in the training dataset and it is shown in Figs. 9(a) and 10(a). TABLE I shows the details of the data set used in our experiment.

<table>
<thead>
<tr>
<th>Epoch</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Batch size</td>
<td>4</td>
</tr>
<tr>
<td>Input / Output image size</td>
<td>256×256</td>
</tr>
<tr>
<td>Input / Output channel</td>
<td>3 ch</td>
</tr>
</tbody>
</table>

B. Static2dynamic Dataset and Implementation

The datasets used for static2dynamic are different with those used for pix2pix. We used pairs of the static expressionless image and its dynamic expression images, which were taken in a dark room at our laboratory. We collected dynamic expression images from 34 individuals (17 male and 17 female individuals). The first frame of each dynamic expression images is used as its corresponding static image (expressionless image). We randomly selected one male and one female as test data. We made two expression datasets: smile expression and anger expression. The images of one female are shown in Fig 8 as examples. The expressionless image, dynamic smile images, dynamic angry images are shown in Figs. 8(a), 8(b) and 8(c), respectively. The images were preprocessed as well as pix2pix described in the previous subsection. First, we extracted a ROI area (128×128 pixels) from each dynamic facial expression image as our datasets. Using small image size for computation is due to the memory of the GPU. Second, we augmented the training dataset in the same way as experiment A resulting in 322560 pairs for training.

The details of the dataset used in the static2dynamic experiment are summarized in Table II.
<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Epoch</td>
<td>900</td>
</tr>
<tr>
<td>Batch size</td>
<td>2</td>
</tr>
<tr>
<td>Input / Output image size</td>
<td>128*128</td>
</tr>
<tr>
<td>Input / Output channel</td>
<td>3ch</td>
</tr>
</tbody>
</table>

C. Experimental results of static expression image generation using pix2pix

The smile and angry facial images generated by the trained generator are shown in Figs. 9 and 10, respectively. Figs 9(a) and 10(a) input images (expressionless images), which are not included in the training dataset. The generated smile image and angry image are shown in Figs. 9(b) and 10(b), respectively. The real smile image and angry image are shown in Figs. 9(c) and 10(c), respectively.

To make a quantitative evaluation, we used “My Pocket” [5], an application service of NTT Communications Corporation, to evaluate the degree of smile. The degree for the generated smile image [Fig. 9(b)] is 97%, while the degree of real smile image [Fig. 9(c)] is only 79%. It means that our generated smile image is better or more natural than the real smile image.

D. Experimental results of dynamic expression generation using static2dynamic.

We generated dynamic facial expression images of smile and angry by using the trained static2dynamic model, which is shown in Fig. 3. Input is an expressionless image which is not included in the training dataset [Fig.11 (a)]. We input an expressionless face image into the model and finally, we got the result of the dynamic facial expression images [Fig. 11 (b): smile, Fig. 11 (c): angry]. The real smile image and angry images of the input are shown in Fig. 11 (d) and Fig. 11 (e). The results were evaluated by users. The subjective evaluations were shown in next sub-section.

E. Subjective evaluations of static expression generation results

The generation results were evaluated by 20 students (participants). Each participant responded to the following two questions with a score of 0 ~ 100. The higher the score, the better the evaluation. We calculated the average and variance of the
scores. The evaluation results for static expression and dynamic expression generations are shown in Tables III and IV, respectively.

Q1. Is the input and output the same person?

Q2. Is the expression natural?

TABLE III. Subjective evaluations of static expression generation results (mean ± standard deviation)

<table>
<thead>
<tr>
<th>Pix2Pix Subjective evaluation</th>
<th>Q1</th>
<th>Q2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expression: Smile</td>
<td>91.75 ± 4.17</td>
<td>75.8 ± 2.2</td>
</tr>
<tr>
<td>Generated expression [Fig.9(b)]</td>
<td>93.3 ± 3.14</td>
<td>68.5 ± 2.7</td>
</tr>
<tr>
<td>Real expression [Fig.9(c)]</td>
<td>95.6 ± 3.23</td>
<td>71.7 ± 4.4</td>
</tr>
<tr>
<td>Expression: Angry</td>
<td>94.85 ± 3.86</td>
<td>86.3 ± 2.7</td>
</tr>
<tr>
<td>Generated expression [Fig.10(b)]</td>
<td>95.6 ± 3.23</td>
<td>71.7 ± 4.4</td>
</tr>
<tr>
<td>Real expression [Fig.10(c)]</td>
<td>96.5 ± 2.7</td>
<td>77.7 ± 11.7</td>
</tr>
</tbody>
</table>

TABLE IV. Subjective evaluations of dynamic expression generation results (mean ± standard deviation)

<table>
<thead>
<tr>
<th>static2dynamic. Subjective evaluations</th>
<th>Q1</th>
<th>Q2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expression: Smile</td>
<td>88.5±11.5</td>
<td>83.2±12.4</td>
</tr>
<tr>
<td>Generated expression [Fig.11(b)]</td>
<td>89.9±12.14</td>
<td>75.2±9.7</td>
</tr>
<tr>
<td>Real expression [Fig.11(d)]</td>
<td>91.6±7.23</td>
<td>67±11.8</td>
</tr>
<tr>
<td>Expression: Angry</td>
<td>85.5±13.4</td>
<td>77.7±11.7</td>
</tr>
<tr>
<td>Generated expression [Fig.11(c)]</td>
<td>91.6±7.23</td>
<td>67±11.8</td>
</tr>
<tr>
<td>Real expression [Fig.11(e)]</td>
<td>96.5±2.7</td>
<td>77.7±11.7</td>
</tr>
</tbody>
</table>

For question 1, the generated static smile, static angry, and dynamic smile have similar scores with the real expression images, and only the score of the generated dynamic angry is lower than that of the real expression. For question 2, all the generated images (static smile, static angry, dynamic smile and dynamic angry) have higher scores than those of the real expression images. The results demonstrated that the proposed method can generate natural dynamic expression facial images.

IV. CONCLUSION

In this paper, we proposed a static-to-dynamic model (static2dynamic) based on a 3D conditional generative adversarial network (3D CGAN), which is an extension of the conventional CGAN. Herein, the dynamic facial expression image is treated as a 3D image. The effectiveness of the proposed method was demonstrated by generating dynamic smile and angry facial expression images.

References


