Personalized Expression Synthesis Using a Hybrid Geometric-Machine Learning Method

Sarra Zaied(✉), Catherine Soladie(✉), and Pierre-Yves Richard(✉)

Institute of Electronics and Telecommunications of Rennes UMR CNRS 6164, Research Team FAST CentraleSupelec, Rennes, France
{sarra.zaied,catherine.soladie,pierre-yves.richard}@centralesupelec.fr

Abstract. Actually, various Geometric and Machine Learning methods are employed to synthesize expressions. The geometric techniques offer high-performance shape deformation but lead to images which are lacking in texture details such as wrinkles and teeth. On the other hand, the machine learning methods (e.g. Generative Adversarial Network GAN) generate photo-realistic expressions and add texture details to the images but the synthesized expressions are not those of the person. In this paper, we propose a hybrid geometric-machine learning approach to synthesize photo-realistic and personalized joy expressions while keeping the identity of the emotion. Our approach combines a geometric technique based on 2D warping method and a generative adversarial network. It aims at benefiting from the advantages of both paradigms and overcoming their own limitations. Moreover, by adding a previous knowledge of the way of smiling of the subject, we personalize the synthesized expressions. Qualitative and quantitative results demonstrate that our person-specific hybrid method can generate personalized joy expressions closer to the ground truth than two generic state-of-the-art approaches.

Keywords: Expression synthesis · Person-specific model · Hybrid method MLS-GAN

1 Introduction

The theory of peripheral emotional feedback states that our emotional experiences are under the retroactive influence of our own expressions [13]. It has been an ongoing subject of debate in psychology since James [9]. Actually, facial expressions become a tool for psychological analysis [11]. The fact that putting on a smile or frown may have an implicit, automatic effect in one's emotional experience holds tremendous potential for clinical remediation in psychiatric disorders. We address this situation head-on, proposing a framework able to channel the psychological mechanism of facial feedback for clinical application

Work funded by ANR REFLETS project.

E. Ricci et al. (Eds.): ICIAP 2019, LNCS 11752, pp. 24–34, 2019.
https://doi.org/10.1007/978-3-030-30645-8_3

to post-traumatic stress disorders (PTSD). Via our framework implemented in a mirror-like device, the observers can see themselves in a gradually more positive way: without their knowing, their reflected face is algorithmically transformed to appear more positive. Using this system, we expect that observers believe the emotional tone of their transformed facial reflection as their own, and align their feelings with the transformation. One more thing, the expressions are highly personal and each person smiles differently. To act positively on the emotional state of a subject while keeping the credibility, we generate a person-specific joy expression basing on previous knowledge of her own way of smiling.

Our model leads to preserve the morphology shape and the identity of the emotion by reproducing the specific way of smiling of each subject. The first contribution is that we personalize the generated expression. As we cannot guess the specific way of smiling of a subject, we need previous knowledge of her way of smiling. The originality is that we first learn a Person-Specific model using one neutral and one smiling face of the person whose expression we want to change. The model is composed of several parameters which are specific to each subject and is used to perform a 2D warping on the neutral face to synthesize a specific and personalize joy expression. The second contribution of our method is that we use a GAN to refine the details in the synthesized images. The originality is that we introduce a hybrid method combining geometric and machine learning tools. The geometric part aims at preserving the identity shape and optimize the distortion made on the image. The GAN offers a realistic facial texture and allows to naturally refine the local-texture details of the synthesized images such as wrinkles and teeth. The last contribution is that our approach can synthesize smile expression with different intensities from a single image.

2 Related Work

Research work on synthesizing photo-realistic facial expressions on real subjects can be divided into two categories: Geometric techniques and machine learning methods. The first category mainly resorts to computer graphic techniques and is based on shape distortion. These methods directly deform the detected landmarks to generate an expressive face [2,12,18,24]. Most of the time, the shape distortion is based on a predefined translation of the landmarks [12,18,24]. The main limitation is that the participants must keep the neutral facial expression and they should not speak or change their head position during the expression generation. These constraints was tackled in the framework of Pablo et al. [2]. The proposed system adapts to the position of the user's face. Yet, the deformations in these works are still based on the same model for all the subjects. Then, these methods generate non-personalized expressions. However, the expressions (e.g. smile) are personal and it occurs in a different way for each person, it may be: straight, curved or elliptical. Finally, the texture refinement in these works are obtained by the moving least squares method [16] (MLS). This method provides a deformation that minimizes the amount of local scaling and optimizes the distortion made on the image in real time. However, it leads to generate images that miss fine details such as wrinkles and teeth.

The second category aims at building generative models [6] to synthesize facial expressions with predefined attributes [3,7,14,21,22]. These models are based on texture refinement. The GANs allow to synthesize different photo-realistic facial expressions [7,21] or to animate a single RGB image [3,14]. These models enable to generate natural and reasonable face expressions (e.g. different smiles) and generate images with fine local details. The deformations are based on a model which is learned on several subjects. So that the synthesized expressions are not the subject's own. Yet, each person has her own way to make expressions. Finally, images generated by such methods sometimes tend to be blurry and of low resolution.

The geometric methods provide a relevant shape deformation but it lacks local details in the generated images. The generative models succeed in adding texture details (wrinkles and teeth) but the generated smiles are not those of the person. To this aim, we propose a hybrid geometric-machine learning method that combines the benefits of the geometric and machine learning methods to synthesize person-specific joy expression. We use a geometric technique to deform the shape of a detected face to appear more positive. For each person, we learn a parametric model according to her own way of smiling using her neutral expression and her smile one. Then we use this model as previous knowledge of the way of smiling to synthesize a personalized joyful expression. As the synthesized images are lacking in details such as teeth and wrinkles, we employ a GAN to refine the texture and to add local fine details to these images. Our framework is able to synthesize person-specific joy expressions with different intensities while preserving the emotion user's identity.

3 Our Hybrid Approach

Our framework aims at generating personalized joy expression as illustrated in Fig. 1. Our algorithm starts with a learning step to build a person-specific model (Sect. 3.1). The model is learned using a neutral and a smile expression of the subject. We use the predefined model and a coefficient d to manipulate the intensity of the generated expression (Sect. 3.2). Finally, we use a GAN to refine the local-texture details on the synthesized images such as wrinkles and teeth (Sect. 3.3).

3.1 Learning a Person-Specific Shape Model

To preserve the identity of the emotion and to generate person-specific expressions, we need previous knowledge of the way of smiling of the subject. To this aim, we lean a person-specific model for each subject using her neutral frame X_n and her smiling frame X_s (Fig. 1: part 1). For the tracking, we use GenFace-Tracker of Dynamixyz [4]. This tracker detects the face and determines precisely the coordinates of 84 landmarks as well as the orientation, scale, and position of the face. A smile is expressed with the rise of the corners of the mouth and cheeks, as well as the lifting of the lower eyelids [5], we selected 10 landmarks

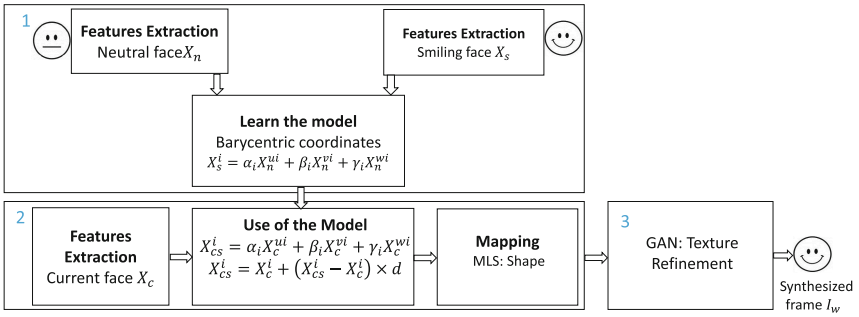

Fig. 1. Our framework is composed of 3 parts. In the first part, we track one neutral X_n and one smiling X_s face of the subject to extract the features which are the landmarks positions. Then, we learn a person-specific model from the detected landmarks of the two faces. In the second part, the learned model is used with each current frame X_c of the subject to generate a current smiling frame X_{cs}. We use the predefined model of the subject and a coefficient d to manipulate the intensity of the generated joy expressions. We employed a 2D warping method MLS [16] to deform the current face. In the third part, we use a generative adversarial network to refine local details on the synthesized frames.

which corresponds to the corners of the mouth and the lower points of the eyes to learn the deformations of the face ($i = \{64, 65, 69, 70, 71, 75, 45, 46, 51, 52\}$). Once the landmarks of the two faces X_n and X_s are aligned by homography, we perform a Delaunay triangulation on the neutral face. Each of the 10 selected landmarks X_s^i is located inside a neutral face triangle ($X_n^{u_i}$, $X_n^{v_i}$, $X_n^{w_i}$). Figure 2 gives an example for the landmark 64. We calculate the barycentric coordinates (α_i, β_i, γ_i) for each of the 10 points X_s^i. These coordinates are the parameters of our person-specific model. The model is composed of 10 vectors with 6 components. These components are the 3 vertices indexes of the triangle in which the landmark is located (u_i, v_i, w_i) and the 3 corresponding barycentric coordinates (α_i, β_i, γ_i). The calculation of X_s^i is formulated as follows:

$$X_s^i = \alpha_i X_n^{u_i} + \beta_i X_n^{v_i} + \gamma_i X_n^{w_i} \tag{1}$$

3.2 Expressions Shape Synthesis with Different Intensities

To generate a joy expression from a current detected image X_c of a subject, We use the learned specific model of this subject as previous knowledge of her way of smiling. We detect the landmarks X_c^i of the current detected face with GenFaceTracker [4]. Having the different coefficients α_i, β_i and γ_i of each of the 10 points of the smiling face, and knowing the coordinates of the triangles (u_i, v_i, w_i) in which are located these points, we determine the positions of the new 10 points of the current smiling face X_{cs}^i (the transformed current face with a smile) using Eq. (2).

$$X_{cs}^i = \alpha_i X_c^{u_i} + \beta_i X_c^{v_i} + \gamma_i X_c^{w_i} \tag{2}$$

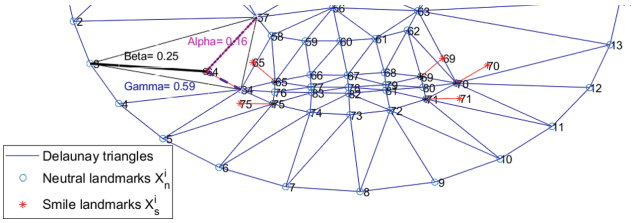

Fig. 2. Positions of the smiling landmarks relative to the neutral landmarks. Each landmark of the smiling face is inside a triangle of the neutral face (X_s^{64} inside the triangle of vertices X_n^{57}, X_n^3, X_n^{64}).

One of the originality's method is that we can generate this expression with different intensities according to a linear model.

$$X_{cs}^i = X_c^i + (X_{cs}^i - X_c^i) \times d \tag{3}$$

Where d is the deformation coefficient. Increasing this coefficient increases the smile intensity and vice versa.

- If $d = 0$, the result is an unchanged face $X_{cs} = X_c$.
- If $d = 1$, the result is an expression of joy X_{cs} that matches the intensity of the one that was learned.

Figure 3 shows an example of the generated joy expressions with different intensities for one subject.

To generate a deformation, that minimizes the amount of local scaling, we applied the Moving Least Squares method MLS [2,16]. The rigid MLS is very effective for image deformation and optimizes the distortion made on the image in real-time. Given the time needed to perform the deformations and apply them, we make a time/esthetic's compromise as in [2]; we apply the algorithm on grids around each eye and around the mouth, not on each pixel of the image.

In this way, for each subject we built a geometric model that is used as a previous knowledge to generate personal joy expressions with different intensities. As shows Fig. 3 the generated joy expression is photo-realistic. However, it misses some details such as teeth and wrinkles.

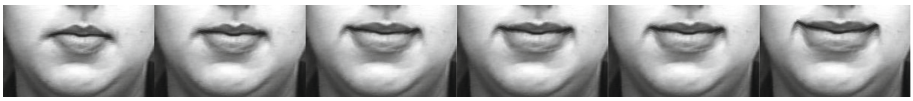

Fig. 3. Generation of 6 joy expressions intensities with coefficient variation from 0 to 1 with step of 0.2.

3.3 Texture Refinement with GAN

To refine the texture details on the synthesized frames (add wrinkles, dimples, and teeth), we use a generative adversarial network [6]. The GAN in [8] achieved impressive results thanks to the used skip connections. It was employed in several researches of facial expressions [14,20,21]. These skip connections between G_{enc} and D_{enc} aims at increasing the resolution of the output. The encoder features are transmitted along these connections with the conditioned information to the decoder. In our framework, we use the same architecture, but we changed the data. Our first originality is that we use the synthesized (warped) images I_{wj} generated with the MLS and the real smiling expressions I_{sj} to train a conditional GAN as Fig. 4 shows. Our second originality is the use of a label vector L_{sj} that is composed of the normalized distance of the lips opening O_j concatenated with one hot vector V_j of 10 values which characterizes the intensity level. This label helps the GAN to correctly add the expression details such as the opening of the mouth, which indicate that it should add teeth. Furthermore, to capture the structural information of the images and achieve more realistic joy expressions, we adopt the feature matching loss term F_m of [15,23]. This function forces the generated image and the real smile image to share the same features. Thus, the generated expression will be closest to the real expression.

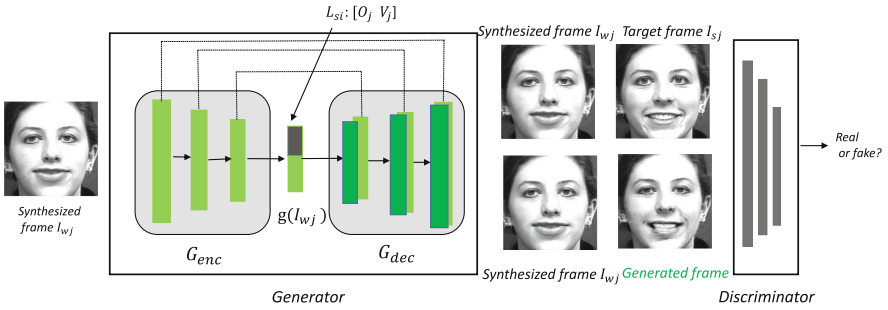

Fig. 4. The GAN is used to refine details on the synthesized geometric images I_{wj}. The synthesized images I_{wj} generated by the geometric step, are fed to the Generator. The encoding representation of the image $g(I_{wj})$ is concatenated with a label vector L_{sj} which is composed of the normalized distance of the lips opening O_j and one hot vector V_j to characterize the intensity. The decoder generates the expressive frame with more details. The discriminator D takes two couple of images; the synthesized frame I_{wj} with the real frame I_{sj} and the synthesized frame I_{wj} with the generated frame to determine if the latter is a real or a fake expression.

4 Experiments and Results

In this section, we present the implementation details on 3 datasets and our qualitative and quantitative results.

Database CK: This database [10] includes 486 sequences from 97 subjects. Each sequence begins with a neutral expression and proceeds to a peak expression. We select the 88 subjects who have smile sequences. Since almost all the sequences are grayscale, we use gray scale images in our experiments.

Database MMI: The database MMI [19] consists of over 2900 videos and high-resolution still images of 75 subjects. It is fully annotated for the presence of AUs in videos. We select the 56 smile videos of 28 subjects which are annotated with AU6 and AU12 (corresponding to smile). Each subject has 2 smiling videos so, we have the opportunity to learn the person-specific model on one video (using a neutral and a smiling face) and test it on another one. Those tests are referred to MMI* in Sect. 4.1.

Database Oulu-CASIA: The database contains 80 subjects [25] with the 6 basic expressions for each subject. We use smile sequences to carry out our experiments. The videos are captured with VIS camera with strong illumination.

The Protocol: With the 3 datasets, we obtain 196 subjects, so we built 196 geometric models. According to the results found in [1,17] concerning the Onset-Apex duration, we choose to synthesize for each subject 10 frames with different intensities. So that we have a total of 1960 synthesized images with the geometric step.

All the synthesized images are normalized, aligned and cropped to 256×256 size to train the GAN. In the training phase, we perform random flipping of the input images to encourage the generalization of the network. We use leave-one-out cross-validation to train the GAN. We use 1950 synthesized smile images of the 195 subjects for the training and 10 images of the remained subject for the test.

For the GAN, we adopt the architecture from [8]. As shown in Fig. 4, the generator G is an auto-encoder U-Net which take as input the synthesized image. The G_{enc} contains 8 convolutional layers. The first one is a simple convolutional layer with a 5×5 kernel and stride 2. The others layers are composed of Leaky ReLU as activation function, convolution with 5×5 stride 2 and a batch normalization. G_{dec} is composed of 8 deconvolutional layers with 5×5 stride 2 and Leaky ReLU. To share the information between the input and the output features, the decoder layers have skip connections with their corresponding layers of G_{enc} (see Fig. 4). Adaptive Moment Estimation optimizer (ADAM) is used to train our model with $\beta = 0.5$, 0.0002 as learning rate and $\lambda = 100$.

In the test phase, we use a new neutral frame of the subject, his learned geometric model and a chosen intensity to synthesize a joyful frame. The frame result is fed to the trained GAN to add the missing details. For the MMI*, as each subject have 2 videos we use one video frames (neutral frame and smile frame) for learning the person-specific model and the second video for the test.

4.1 Quantitative Results

The metric we use to check the performances of our method is the angles between the ground truth smile trajectory and the trajectories of the generated smile with

the method. This allows us to evaluate if the generated smile is that of the person or not and it permits to compare the slope of the real smile and the generated one. As shows Fig. 5, we define the trajectory as the 2D displacement of the mouth landmarks during the smile across the frames. To determine the angles, we use the linear regression $Y = a_{GT}X + b_{GT}$ of the ground truth trajectory. We assume that $Y = aX + b$ is the defined trajectory of the landmarks generated with one method. The angle is determined by:

$$\theta = \tan^{-1} \left| \frac{a_{GT} - a}{1 + a_{GT}a} \right| \tag{4}$$

Table 1 shows the statistical results for the landmark located on the left corner of the mouth. The same study was performed for the 10 landmarks involved in the joy expression. The results show that our method generates trajectories which are closer to the ground truth than the generic methods [2] and [21] ($\overline{\theta}$ closer to 0). In addition, we can consider that our method is more stable than the other two methods because of the low value of σ_θ. We observe that the results with MMI* are less good than the results with MMI because we use a second video of the person to test the model learned with another video of that person. But, we always stay better than the other two methods [2] and [21].

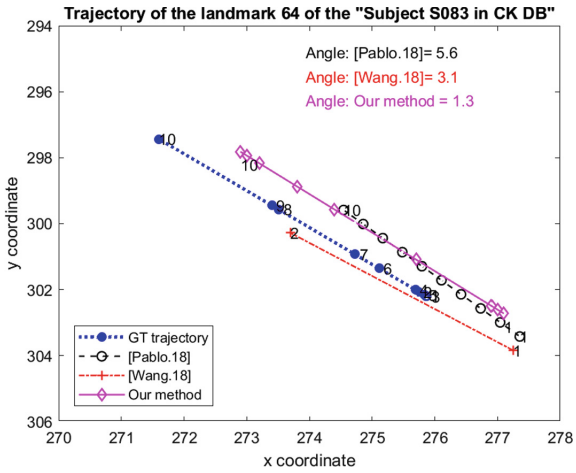

Fig. 5. The ground truth and the generated trajectories of the landmark located at the left corner of the mouth. We can notice that the angle between the GT trajectory and our method trajectory is smaller than the angle calculated with the two generic methods [2] and [21].

We measured the latency of the overall algorithm. Our tests performed on an Intel processor Core i7 at 3.30 GHz with an NVIDIA geforce GTX 1070. The mean time to process a single frame is of 65 ms. It is suitable for real time applications (15 fps) such as the mirror in our use case.

Table 1. Mean and standard deviation of angles calculated with the 3 methods on the 3 databases for the landmark of the left corner of the mouth for all the subjects. With MMI*, we learn a person-specific model with one video of the person and we test it with a second video.

Method	Database			
	CK	Oulu-CASIA	MMI	MMI*
[Pablo.18]	$\overline{\theta} = 12.10$	$\overline{\theta} = 17.81$	$\overline{\theta} = 12$	$\overline{\theta} = 12$
	$\sigma_\theta = 9.73$	$\sigma_\theta = 15.94$	$\sigma_\theta = 13.70$	$\sigma_\theta = 13.70$
[Wang.18]	$\overline{\theta} = 16.16$	$\overline{\theta} = 26$	$\overline{\theta} = 15.26$	$\overline{\theta} = 15.26$
	$\sigma_\theta = 12.24$	$\sigma_\theta = 18.21$	$\sigma_\theta = 12.66$	$\sigma_\theta = 12.66$
Our	$\overline{\theta} = \mathbf{7.65}$	$\overline{\theta} = \mathbf{6.85}$	$\overline{\theta} = \mathbf{5.83}$	$\overline{\theta} = \mathbf{9.20}$
	$\sigma_\theta = \mathbf{8.25}$	$\sigma_\theta = \mathbf{7.26}$	$\sigma_\theta = \mathbf{7.01}$	$\sigma_\theta = \mathbf{10.78}$

4.2 Qualitative Results

Qualitative results confirm that our method gives better results than [2] and [21]. Figure 6 illustrates the results[1] obtained for 2 intensities for 4 subjects with the 3 methods.

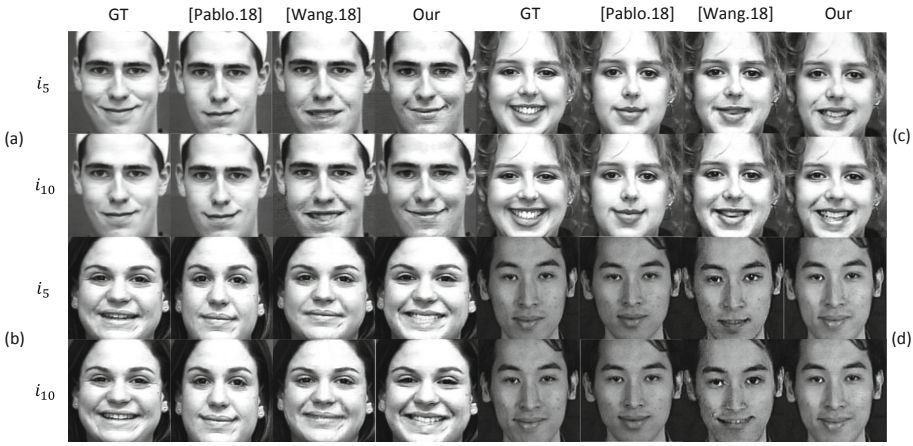

Fig. 6. The Ground Truth smile (GT) and the result frames of 4 subjects from the 2 databases (MMI and CK) with two intensities for each subject.

Concerning the shape, we observe that with the geometric method of Pablo et al. [2], the corner of the lips is systematically raised (steep slope) for all the subjects. We notice that the subjects (a) and (b) have a flat smile (low slope)

[1] More results are available on https://drive.google.com/file/d/1hY15DNrrYNjxnf9JmVFJhWUYhteTYnQa/view?usp=sharing.

in reality while the smiles generated with [2] is growing and not realistic. The GAN [21] generate different smiles but we can perceive that are not those of the subjects. For the subject(c), the real smile is opened but the GAN generate a tight smile for all the subjects.

Concerning the texture, We notice that with Pablo et al. all the generated smiles are without teeth whatever the texture shape of the smile. For the subjects (a) and (d), the real smile is without teeth but the GAN of [21] generate the teeth for these subjects. On the contrary, for the subject (b) the real smile has occurred with teeth but the GAN generate it without. Then, the GAN generate realistic joy expression but not those of the person.

The visual fidelity shows that our method generates the closest smile shape to the ground truth and maintain the same texture of the real smile for all the subjects.

5 Conclusion

In this paper, we proposed a hybrid approach aiming at learning a person-specific model for each person and transforming a captured face to appear more joyful. Our method generate for each subject a photo-realistic joy expression according to her own expression. Qualitative and quantitative results show that the synthesized joy expression with our method is closer to the ground truth than the expression generated with two generic methods. One of our research directions is make psychiatric experiments with our tool to see if we can act on the PTSD patients emotions.

References

1. Ambadar, Z., Cohn, J.F., Reed, L.I.: All smiles are not created equal: morphology and timing of smiles perceived as amused, polite, and embarrassed/nervous. J. Nonverbal Behav. **33**(1), 17–34 (2009)
2. Arias, P., Soladie, C., Bouafif, O., Robel, A., Seguier, R., Aucouturier, J.J.: Realistic transformation of facial and vocal smiles in real-time audiovisual streams. IEEE Trans. Affect. Comput. (2018)
3. Ding, H., Sricharan, K., Chellappa, R.: ExprGAN: facial expression editing with controllable expression intensity. In: AAAI (2018)
4. Dynamixyz: Genfacetracker: person-independent real-time face tracker (2017). http://www.dynamixyz.com
5. Ekman, P., Friesen, W.V.: Facial Action Coding System: Investigatoris Guide. Consulting Psychologists Press (1978)
6. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)
7. Huang, Y., Khan, S.: A generative approach for dynamically varying photorealistic facial expressions in human-agent interactions. In: Proceedings of the 2018 on International Conference on Multimodal Interaction, pp. 437–445. ACM (2018)
8. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017

9. James, W.: What is an emotion? Mind **9**(34), 188–205 (1884)
10. Kanade, T., Cohn, J.F., Tian, Y.: Comprehensive database for facial expression analysis. In: Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition, pp. 46–53. IEEE (2000)
11. Leo, M., et al.: Computational assessment of facial expression production in ASD children. Sensors **18**(11), 3993 (2018)
12. Nakazato, N., Yoshida, S., Sakurai, S., Narumi, T., Tanikawa, T., Hirose, M.: Smart face: enhancing creativity during video conferences using real-time facial deformation. In: Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing, pp. 75–83. ACM (2014)
13. Niedenthal, P.M., Mermillod, M., Maringer, M., Hess, U.: The simulation of smiles (SIMS) model: embodied simulation and the meaning of facial expression. Behav. Brain Sci. **33**(6), 417–433 (2010)
14. Olszewski, K., et al.: Realistic dynamic facial textures from a single image using GANs. In: IEEE International Conference on Computer Vision (ICCV), pp. 5429–5438 (2017)
15. Salimans, T., Goodfellow, I.J., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training GANs. CoRR abs/1606.03498 (2016). http://arxiv.org/abs/1606.03498
16. Schaefer, S., McPhail, T., Warren, J.: Image deformation using moving least squares. In: ACM Transactions on Graphics (TOG), vol. 25, pp. 533–540. ACM (2006)
17. Schmidt, K.L., Bhattacharya, S., Denlinger, R.: Comparison of deliberate and spontaneous facial movement in smiles and eyebrow raises. J. Nonverbal Behav. **33**(1), 35–45 (2009)
18. Suzuki, K., et al.: FaceShare: mirroring with pseudo-smile enriches video chat communications. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, pp. 5313–5317. ACM (2017)
19. Valstar, M., Pantic, M.: Induced disgust, happiness and surprise: an addition to the MMI facial expression database. In: Proceedings of the 3rd International Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect, p. 65 (2010)
20. Wang, T., Liu, M., Zhu, J., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional GANs. CoRR abs/1711.11585 (2017). http://arxiv.org/abs/1711.11585
21. Wang, X., Li, W., Mu, G., Huang, D., Wang, Y.: Facial expression synthesis by u-net conditional generative adversarial networks. In: Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval, pp. 283–290. ACM (2018)
22. Wu, X., Xu, K., Hall, P.: A survey of image synthesis and editing with generative adversarial networks. Tsinghua Sci. Technol. **22**(6), 660–674 (2017)
23. Xu, X., Sun, D., Pan, J., Zhang, Y., Pfister, H., Yang, M.H.: Learning to super-resolve blurry face and text images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition CVPR, pp. 251–260 (2017)
24. Yoshida, S., Tanikawa, T., Sakurai, S., Hirose, M., Narumi, T.: Manipulation of an emotional experience by real-time deformed facial feedback. In: Proceedings of the 4th Augmented Human International Conference, pp. 35–42. ACM (2013)
25. Zhao, G., Huang, X., Taini, M., Li, S.Z., PietikäInen, M.: Facial expression recognition from near-infrared videos. Image Vis. Comput. **29**(9), 607–619 (2011)