05 Jun 2018

Identity-adaptive Facial Expression Recognition Through Expression Regeneration Using Conditional Generative Adversarial Networks

Huiyuan Yang
Missouri University of Science and Technology, hyang@mst.edu

Zheng Zhang

Lijun Yin

Recommended Citation

H. Yang et al., "Identity-adaptive Facial Expression Recognition Through Expression Regeneration Using Conditional Generative Adversarial Networks," *Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018*, pp. 294 - 301, Institute of Electrical and Electronics Engineers, Jun 2018.
The definitive version is available at https://doi.org/10.1109/FG.2018.00050

Identity-Adaptive Facial Expression Recognition Through Expression Regeneration Using Conditional Generative Adversarial Networks

Huiyuan Yang, Zheng Zhang and Lijun Yin
Department of Computer Science
State University of New York at Binghamton, USA
{*hyang51, zzhang27*}*@binghamton.edu; lijun@cs.binghamton.edu*

***Abstract*—Subject variation is a challenging issue for facial expression recognition, especially when handling unseen subjects with small-scale lableled facial expression databases. Although transfer learning has been widely used to tackle the problem, the performance degrades on new data. In this paper, we present a novel approach (so-called *IA-gen*) to alleviate the issue of subject variations by regenerating expressions from any input facial images. First of all, we train conditional generative models to generate six prototypic facial expressions from any given query face image while keeping the identity related information unchanged. Generative Adversarial Networks are employed to train the conditional generative models, and each of them is designed to generate one of the prototypic facial expression images. Second, a regular CNN (FER-Net) is fine-tuned for expression classification. After the corresponding prototypic facial expressions are regenerated from each facial image, we output the last FC layer of FER-Net as features for both the input image and the generated images. Based on the minimum distance between the input image and the generated expression images in the feature space, the input image is classified as one of the prototypic expressions consequently. Our proposed method can not only alleviate the influence of inter-subject variations, but will also be flexible enough to integrate with any other FER CNNs for person-independent facial expression recognition. Our method has been evaluated on CK+, Oulu-CASIA, BU-3DFE and BU-4DFE databases, and the results demonstrate the effectiveness of our proposed method.**

***Keywords*-FER; GAN; Identity-adaptive; CNN;**

I. Introduction

In the past decades, research on automatic facial expression recognition has been conducted through classifying classic prototypic facial expressions (*i.e.*, anger, disgust, fear, happiness, sadness, surprise) from static images or dynamic image sequences. Although significant progresses have been made in recent years, the challenge still remains in real-world scenarios with respect to various poses, illumination, occlusion, and in particular, the identity related expression variations. The inter-subject variations come from a variety of identity components, including age, gender, race and person-specific characteristics [1]. Compared with the VGG face dataset [2] and Imagenet [3], the current public expression datasets are limited in size, making it more difficult to deal with identity-related variations.

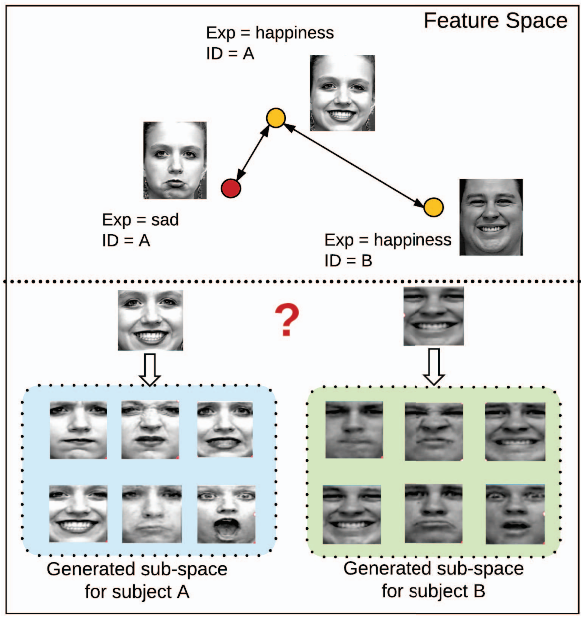

Figure 1. The upper part is the illustration for features of different subjects with a variety of expressions in the feature space. The lower part shows the generated identity-adaptive sub-space for each subject.

Recently, several approaches have been proposed by focusing on improving performance on person-independent facial expression recognition [4] [1] [5]. Transfer learning is one of the mostly used methods, which fine-tunes a network that has been pre-trained on a large dataset *i.e. VGG [2], FaceNet [6] and FER-2013 [7]*, to a relatively small facial expression dataset. Other methods include: using a deeper network with more training data [8]; learning a person-specific model [9] and adding extra constraints for identity-related variations in FER task. Although those efforts have alleviated the problem to a certain degree, the challenge still remains unsolved.

For facial expression recognition task, the same expressions are expected to have a smaller distance between each other in the feature space than others of different expres-

978-1-5386-2335-0/18/$31.00 ©2018 IEEE
DOI 10.1109/FG.2018.00050

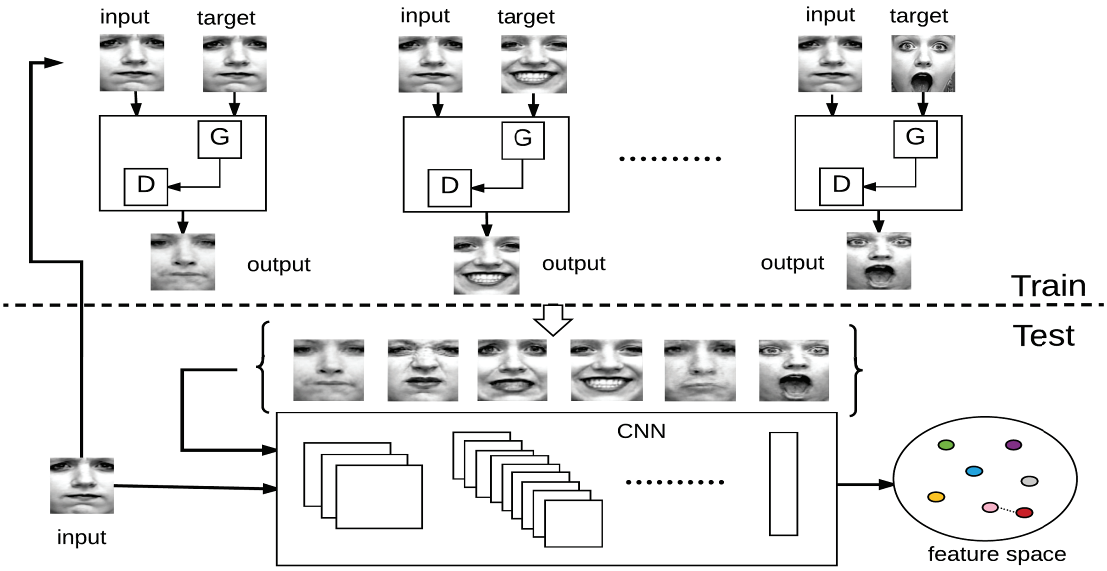

Figure 2. Framework of our proposed method (**IA-gen**)

sions. As illustrated in Fig. 1, $I_{(happiness,A)}$ and $I_{(sad,A)}$ are the same subject A showing different expressions, while $I_{(happiness,A)}$ and $I_{(happiness,B)}$ are different subjects showing the same expressions. However, due to high inter-subject variations, the distance between $I_{(happiness,A)}$ and $I_{(happiness,B)}$ is larger than $I_{(happiness,A)}$ and $I_{(sad,A)}$.

Variations are greater related to identity compared to variations related to expression. This partially explains why most current methods could degrade on unseen subjects. As in the lower part of Fig. 1, we generate two sub-spaces for subject A and subject B respectively, each of them contains the generated six basic expression images of the query image. Only the same identity information exists in each sub-space, thus allowing for expression classification to be more reliable. This motivated us to design a new scheme by generating a kind of sub-space for facial expression classification while aiming each subject individually. This identity-adaptive FER approach is robust to identity variations, achieving the goal of person-independent facial expression recognition.

Generative Adversarial Networks (GAN) [10], an effective training method for training generative models, was first developed by Goodfellow et al in 2014. The GAN framework plays an adversarial game with two players, a *generator* and *discriminator*. The discriminator is designed to distinguish between samples from the generator and samples from the training data. On the other hand, the generator learns to output samples that can maximally confuse the discriminator. The *conditional* Generative Adversarial Networks (cGAN) [11], an extension of the basic GAN model, enables the generative model to learn different contextual information by adding extra conditional variations to the input. Another extension is the combination of GAN and CNN [12], which has facilitated a number of interesting applications, *e.g*, object attributes manipulation by vector arithmetic [12], face generation [13], and object reconstruction from edge maps [14].

The potential application of cGAN in face generation, as well as the current limitation of facial expression recognition, motivated us to develop an identity-adaptive structure to address the issue of high inter-subject variations through regenerating six basic facial expression images of a same subject. Unlike the previous method [1] that tried to split the facial image feature into two parts: expression-related and identity-related. However, our method is based on the assumption that individual facial expression is dependent on the individual's identity, *i.e. gender, age, and race with various expression styles*, whereby expression-related features and identity-related features are partially overlapped, and not separable. By regenerating six facial expression images of a subject given a face image of the individual, our approach limits the feature comparison in a single identity sub-space, allowing us to further compare the features merely caused by expression variations of the subject. The feature space is adaptive to a single identity without involving other individuals, thus identity variations will not be an issue for facial expression classification. This enables the facial expression recognition to work in a person-independent manner. The main contribution of this paper is two-fold:

- To our knowledge, this is the first work to address the facial expression recognition by expression re-generation through the exploration of GANs and the application of cGANs.

- The proposed method can effectively deal with the issue of subject variations, because the feature sub-space is generated for a single subject with no affiliation with the other subjects. This ”adaptive” property allows FER to work person-independently. In addition, the proposed method can be easily integrated into regular networks for performance improvement.

To validate the effectiveness of our proposed method, we conducted four experiments on four public databases: CK+ [15], Oulu-CASIA [16], BU3DFE [17] and BU-4DFE [18]. The results show that our proposed method achieved superior performance compared to the state-of-the-art methods.

II. Related Works

Facial expressions are varied from person to person with respect to age, race, gender, and cultural background in terms of their appearances and styles of facial actions. Such a **subject variation issue** could cause performance degradation in facial expression recognition, especially on unseen subjects. There are existing approaches developed for focusing on improving person-independent facial expression recognition. Chen et al. [9] attempted to learn a person-specific model through transfer learning. Ding et al. [5] proposed a so-called FaceNet2ExpNet structure for facial expression recognition on relatively small datasets, which used the pre-trained model as supervision rather than as a source of some initial parameter values. Their method consists of two stages: first, only the convolutional layers were trained to generate similar intermediate features to those of the pre-trained model; second, the whole deep model was trained with the label information. The experiments showed that their method worked well on relatively small expression datasets. Zhao et al. [4] achieved invariance to expression intensity by considering both peak expressions and weak expressions to train a network. The invariance was achieved by a peak gradient suppression learning algorithm, which drove the intermediate features of weak expressions towards those of peak expressions. Meng et al. [1] proposed an identity-aware deep model for facial expression recognition, which was capable of learning the features that were invariant to both expression and identity-related variations, by utilizing an expression-sensitive contrastive loss function.

Generative Adversarial Networks (GANs) have been vigorously studied in recent years. It was first proposed by Goodfellow et al. [10] known as the *generative adversarial net* (GAN), which applied the minimax game to two players, *i.e.*, *generator* (G) and *discriminator* (D) for recovering the distribution of the training data by G while keeping D to $\frac{1}{2}$. Radford et al. [12] had successfully scaled up GANs using CNNs to model images, and showed good result with the trained discriminators for image classification tasks, as well as the interesting vector arithmetic properties of generated samples of generator. Gauthier [13] extended the generative adversarial net framework by adding a conditional information, which could deterministically control the output of the generator, and showed how to generate faces with specific attributes from nothing but random noise and conditional information. Isola et al. [14] utilized conditional adversarial networks for image-to-image translation, and showed that the conditional adversarial networks were applicable for a wide variety of applications, *i.e. synthesizing images from labels, reconstructing objects from edge maps and colorizing images.*

Inspired by the above-mentioned prior works, especially by the solution of image-to-image translation with conditional generative adversarial networks [14], we propose to explicitly train a set of expression-specific models (*e.g.*, six prototypic expression models) to generate the corresponding expressions of each subject for a given image of the subject. By doing so, we can utilize expression features oriented by each individual adaptively in the corresponding feature sub-space even with the identity-related variations occurred from person to person, thus leading to the FER improvement with unseen subjects.

III. Proposed Method

Our proposed method is based on the regeneration of six basic facial expressions, which are adaptive to each identity of individual, we call it the Identity-Adaptive Generation method (a.k.a. **IA-gen**). As shown in Fig.2, the **IA-gen** framework consists of two parts. The upper part is aimed to generate six basic facial expression images of the same subject for any query image using six cGANs, and each of them is designed to generate one expression respectively. The lower part is the facial expression recognition module. A pre-trained CNN is first fine-tuned on the database, and then the last fully connected layer is used as features for both the query image and regenerated images. The query image is labeled as one of the six basic expressions based on a minimum distance in feature space.

A. GANs for Facial Expression Regeneration

GANs have been successfully used to generate images [14] [13] [12]. Note that CNN based methods could also be used for image generation. However, they use a mean squared error (MSE) based solution, resulting in overly smooth of the generated image. However, GANs force the regeneration towards the natural image manifold and produce more perceptually natural results [19]. Therefore, we apply the GANs for facial expression image regeneration.

The generator G is trained to produce outputs that cannot be distinguished from ”real” image pairs by discriminator D, which is adversarially trained to detect ”fake” image pairs. The training procedure is illustrated in Fig. 3, where G generates an output for any input image, and the image pairs $< I_{input}, I_{output} >$ are constructed as negative examples

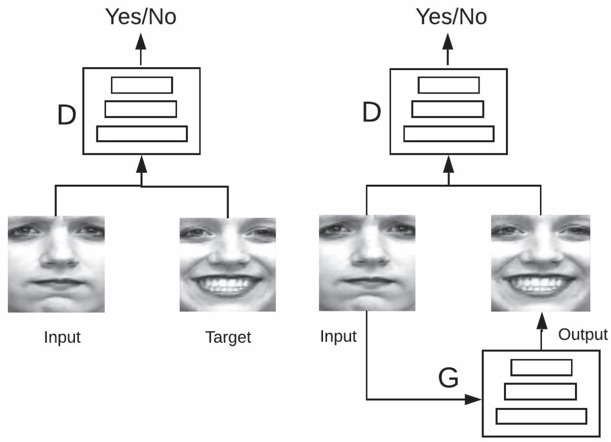

Figure 3. Training a GAN to generate facial expression image. The discriminator (D) learns to distinguish between the [input, target] and [input, output] pairs, while the generator (G) learns not only to fool the discriminator, but also to be close to the ground truth (target image).

for the D, while image pairs $< I_{input}, I_{target} >$ from the training dataset are also constructed as positive examples.

The objective for discriminator can be expressed as:

$$L_{cGAN}(D) = \frac{1}{N}\sum_{i=1}^{N}\Big\{ logD(I_{input}^{i}, I_{target}^{i}) + log\big[1 - D\big(I_{input}^{i}, G(I_{input}^{i})\big)\big]\Big\} \quad (1)$$

where N is the total number of training image pairs $< I_{input}, I_{target} >$. To against an adversarial D, G is used to not only fool the discriminator, but also to generate an output as similar to the target image as possible. The loss function for the generator is defined as follow:

$$L_{cGAN}(G) = \frac{1}{N}\sum_{i=1}^{N}\Big\{ L_{adv} + \alpha \cdot L_{cont}\Big\} \quad (2)$$

$$L_{adv} = -log\big[D\big(I_{input}, G(I_{input})\big)\big] \quad (3)$$

Here, L_{adv} is the adversarial loss, as defined in Eq.3, $D\big(I_{input}, G(I_{input})\big)$ is the probability that the regenerated image pairs $< I_{input}, I_{output} >$ are recognized as training example. L_{cont} is the content loss between the regenerated I_{output} and training example I_{target}. The most widely used loss function is the pixel-wise MSE loss, which is calculated as:

$$L_{cont}^{MSE} = \frac{1}{c^2WH}\sum_{x=1}^{cW}\sum_{y=1}^{cH}\Big\{ I_{target}(x,y) - I_{output}(x,y)\Big\}^2 \quad (4)$$

Here, *c, W, H* are the channel, width and height of the image, respectively. However, MSE based solution often overly smooths textures, which results in the lack of high frequency content, and generates perceptually unsatisfying results [19]. In our experiment, we use L_1 loss rather than L_2 loss, because the L_1 loss can reduce the effect of over-smoothing on the generated image [14]. Another choice for content loss is perceptual similarity that measure high-level perceptual differences between images [20]. The perceptual loss is defined on a loss network ϕ which is pre-trained for image classification, e.g., VGG network [2]. Rather than penalizing the pixel differences between I_{output} and I_{target}, the perceptual loss allows to have similar feature representations to be computed in ϕ. Let $\phi_j(I)$ be the *j-th* convolutional layer with a shape of $C_j \times H_j \times W_j$; then the perceptual loss is the euclidean distance between feature representations:

$$L_{cont}^{pep} = \frac{1}{C_j^2 H_j W_j}\sum_{x=1}^{C_jW_j}\sum_{y=1}^{C_jH_j}\Big\{ \phi_{x,y}(I_{target}) - \phi_{x,y}(I_{output})\Big\}^2 \quad (5)$$

The final loss function for the generator G can be written as:

$$L_{cGAN}(G) = \frac{1}{N}\sum_{i=1}^{N}\Big\{ L_{adv} + \alpha \cdot L_{cont}^{MSE} + \beta \cdot L_{cont}^{pep}\Big\} \quad (6)$$

and the final optimization target is to solve the adversarial min-max problem:

$$G^* = arg \min_G \max_D \big\{ L_{cGAN}(D) + \lambda L_{cGAN}(G)\big\} \quad (7)$$

The complete architecture of GAN follows the structure used in [14]. Specifically, the generator G is a deconvolutional neural network [21], which contains six Encoders with output channels {64, 128, 256, 512, 512, 512} and six Decoders with output channels {512, 512, 256, 128, 64, 1}. Each layer is followed by a non-linear activation function (ReLU) and batch normalization (BN) layer. The stride size is set as 2 to avoid max-pooling operation, and the filter size is set as 4×4 for all layers. Skip connections (U-net [22]) are also used here, which help pass low-level information shared between the input and output image directly across the net. The discriminator D is a regular convolutional neural network with {64, 128, 256, 512, 1} output channels.

B. A CNN for Facial Expression Recognition

Due to large variations across subjects, facial expression recognition performance degrades on unseen subjects in real world scenarios. To cope with this problem, IA-gen is first used to construct a sub-space with six expressions of the same subject, and then a regular convolutional neural networks (FER-net) is used for facial expression recognition. The last fully connected layer of FER-net is used as features for both the input image and the generated images, and the input image is labeled as one of the six basic facial expressions based on a distance function in the feature space:

$$Predict = arg \min_i ||Feat(I_{input}) - Feat(I_i)|| \quad (8)$$

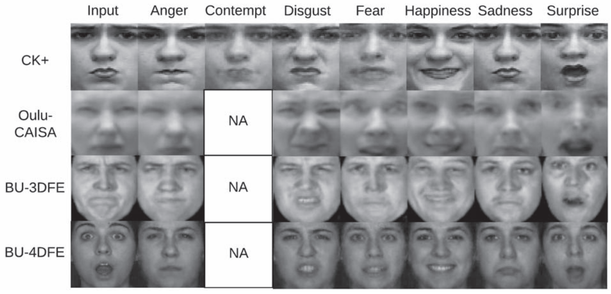

Figure 4. The generated six facial images for the input image in four databases.

Here, $Feat(\cdot)$ is a feature extraction function. Any method, except the FER-net used here, that can be used to extract the facial expression related features works as well.

IV. Experiments

We apply the proposed **IA-gen** approach to the task of facial expression recognition on four publicly available facial expression databases: extended CK+ [15], Oulu-CASIA [16], BU-3DFE [17] and BU-4DFE [18].

A. Implementation

We apply the tree-structured deformable models (TSM) [23] for face detection and landmarks localization. The faces are then cropped and resized to 70 × 70. The labeled facial expression database is quite small, thus we utilize conventional data augmentation method to generate more training data, where each image is rotated by degree $[-15^o, -12^o, -9^o, -6^o, -3^o, 0^o, 3^o, 6^o, 9^o, 12^o, 15^o]$ respectively. Five 64 × 64 patches are cropped out from five locations of each image (*center and four corners, respectively*), then each patch is flipped horizontally, thus resulting in an augmented dataset which is 110 times larger than the original one. During testing, neither the rotation nor the flipping operation is used on the input image.

The cGAN models are pre-trained on BU-4DFE database, and then fine-tuned on other databases. In order to have a similar baseline as [5], the VGG [2] model is used as baseline in Oulu-CAISA database. All the others use the FER-Net, which is fine-tuned from a pre-trained CNN model on the Facial Expression Recognition (FER-2013) database [7]. The batch size is 100, the momentum is fixed to be 0.9, and the dropout is set to 0.5 during training, and 1.0 during testing. The initial learning rate is set to 0.001, and is decreased by 0.8 for 20 epochs. α, β and λ are set to 100, 10 and 1 respectively. Both cGANs and FER-Net are implemented using tensorflow [24].

B. Facial Expression Generation

Fig.4 illustrates the generated facial expression images by cGAN. The first column (from top to bottom) represents the input images from CK+ [15], Oulu-CASIA [16], BU3DFE [17] and BU-4DFE [18] respectively, and the rest of the columns show the generated facial images with different expressions. As illustrated in Fig.4, the facial expression regeneration part is capable of generating different facial expression images while keep the identity-related information visually unchanged.

Table I
Average accuracy on the CK+ database for seven classes.

Method	Setting	Accuracy
LBP-TOP [25]	sequence-based	88.99
HOG 3D [26]	sequence-based	91.44
3DCNN [27]	sequence-based	85.9
STM-Explet [28]	sequence-based	94.19
IACNN [1]	image-based	**95.37**
DTAGN [29]	sequence-based	97.25
CNN (baseline)	image-based	89.50
Ours	image-based	**96.57**

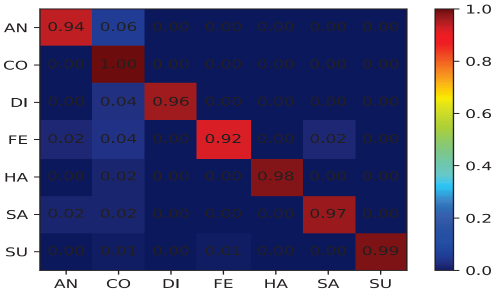

Figure 5. Confusion matrix on the CK+ database.

C. Evaluation on CK+

The Extended Cohn-Kanade database (CK+) [15] is widely used for evaluating facial expression recognition. It contains 593 video sequences recorded from 123 subjects, each of which displayed one of seven expressions, from neutral to peak expression. The label information is only provided for the last frame of each sequence. Following the general procedure, we use the last three frames of each sequence with the provided label, which results in 981 images. The images are further split into 10 folds by ID in ascending order, so the subjects in any two subsets are mutually exclusive. Both the proposed method and CNN based baseline are trained and tested on static images. The reported results are the average of the 8 runs. As shown in Table I, the performance of our proposed method outperforms the CNN based baseline in terms of the average accuracy of seven expressions. The proposed method is also compared to the state-of-the-art methods evaluated on the CK+ database. Among them. IACNN [1], CNN baseline and our proposed method are image-based methods, while others are sequence-based, and use the temporal information. In contrast, our method IA-gen, which is more suitable for ap-

Table II
AVERAGE ACCURACY ON THE OULU-CASIA DATABASE.

Method	Setting	Accuracy
LBP-TOP [25]	sequence-based	68.13
HOG 3D [26]	sequence-based	70.63
STM-Explet [28]	sequence-based	74.59
Atlases [30]	sequence-based	75.52
DTAGN-Joint [29]	sequence-based	81.46
FN2EN [5]	image-based	87.71
PPDN [4](peak expression)	image-based	84.59
VGG (baseline)	image-based	82.33
Ours	image-based	**88.92**

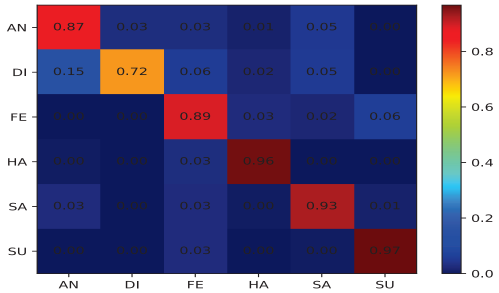

Figure 6. Confusion matrix on the Oulu-CASIA database.

Table III
AVERAGE ACCURACY ON THE BU-3DFE DATABASE.

Method	Setting	Accuracy
Wang et al. [31]	3D	61.79
Berretti et al. [32]	3D	77.54
Yang et al. [33]	3D	**84.80**
Lopes [34]	image-based	72.89
CNN (baseline)	image-based	73.2
Ours	image-based	**76.83**

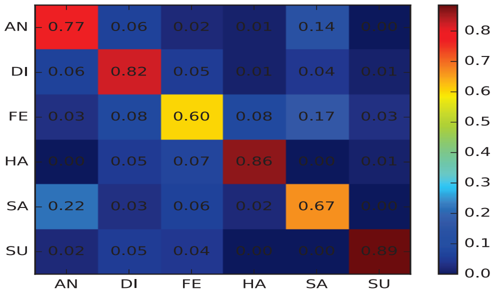

Figure 7. Confusion matrix on the BU-3DFE database.

plications where videos or image sequence is not available, achieves comparable result, with 96.57% for seven classes. Aimed at solving the identity-related issues, IACNN [1] tried to split the whole face image feature into two parts: identity-related and expression-related features, but this method may not generalize well for all six basic facial expressions. For example, people may show something in common when smiling, but often individuals have unique ways to express other emotions. This difference can be related to age, gender, race, and even the background of education. Instead of trying to split the identity-related information (which, sometimes, is not separable), our proposed IA-gen adaptively regenerates a sub-space that contains six basic facial expressions of the same subject, so the facial expression recognition is limited in a single-identity subspace, and that's why our proposed method shows higher performance than IACNN [1].

Fig. 5 is the confusion matrix, from which we can see that, our proposed method performs well in recognizing *contempt* and *surprise*, while showing lowest recognition accuracy in *fear*, which is confused mainly with *contempt* and *anger*.

D. Evaluation on Oulu-CASIA

The Oulu-CASIA database [16] contains two subsets: VIS and NIR. Here, we only use the VIS subset, in which all videos were captured by VIS camera. The Oulu-CASIA VIS has 480 video sequences taken from 80 subjects and six expressions under dark, strong, weak illumination conditions. In this experiment, only videos that were captured under strong condition are used. As a general procedure, the last three frames of each sequence with the provided label are used, which result in 1440 images. Similar to experiment settings in [4], a 10-fold subject independent cross validation is performed.

The 82.33% baseline is reported by fine-tuning the VGG [2] network on the dataset. With the introduction of **IA-gen** approach, the classification performance is increased to 88.92%.We also compare with the state-of-the-art methods, as summarized in Table II. Among them, FN2EN [5], PPDN [4] and our proposed method are image-based, while the others are sequence-based methods, which utilize the temporal-spatio information for facial expression recognition.

In the confusion matrix in Fig.6, the *happiness* and *surprise* expressions have the highest recognition rate, while *disgust* shows the lowest recognition accuracy, and is mainly confused with *anger*.

E. Evaluation on BU-3DFE

The BU-3DFE database [17] contains 100 subjects, ranging in age from 18 to 70 years old with a variety of race. Each subject displays six basic expressions with four levels of intensity and a neutral expression, which results in a total of 2,500 3D facial expression models and texture images. In this experiment, we only use the texture images, and high intensity expressions (the last two frames). A 10-fold cross-validation is performed, and the split is subject independent. Table III shows the average result of 10 runs on the BU-3DFE database. Unlike the CK+ and Oulu-CASIA databases, the BU-3DFE is more challenging as it

Table IV
AVERAGE ACCURACY ON THE BU-4DFE DATABASE.

Method	Setting	Accuracy
Dapogny et al. [35]	sequence-based	75.8
LSH-CORF. [36]	sequence-based	77.1
PCRF [37]	image-based	76.1
László et al. [38]	3D	78.2
Pan et al. [39]	image-based	80.89
CNN (baseline)	image-based	76.45
Ours	image-based	**89.55**

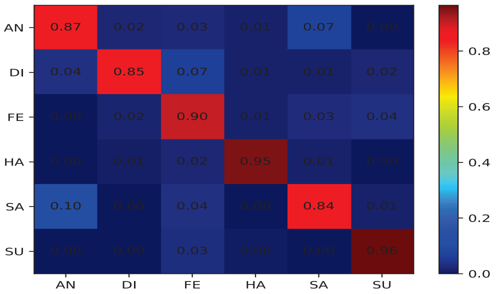

Figure 8. Confusion matrix on the BU-4DFE database.

has a wider variety of ethnicities and a larger range of ages while the data size is relatively small. Note that [33] has higher performance due to the use of geometric feature of the 3D shape model. However, our proposed method improves the baseline by 3.6%, and also outperforms the image-based method [34].

Fig. 7 shows the confusion matrix, where the expressions (surprise, happiness, and disgust) are classified better than the other expressions.

F. Evaluation on BU-4DFE

The BU-4DFE [18] contains 606 facial expression sequences captured from 101 subjects. Each sequence shows one of the six basic facial expressions starting from neutral to peak expression, and ending with neutral again. Seven frames around the peak expression are collected with the provided sequence label, which results in $4272(101 \times 6 \times 7)$ images. Similar to the CK+ database setting, a 10-fold subject-independent cross validation is performed. Table IV reports the average accuracy of 10 runs on the BU-4DFE database for recognizing six expressions. As shown, our proposed method has a 13.1% improvement over the CNN baseline, and achieves the highest accuracy compared to the stat-of-the-art methods, including sequence-based methods [35] [36], image-based methods [37] [39] and 3D model based method [38]. The confusion matrix in Fig.8 shows that *happiness* and *surprise* are the easiest expressions to recognize, while *sadness* shows relatively low recognition rate, which is mainly confused with *anger*.

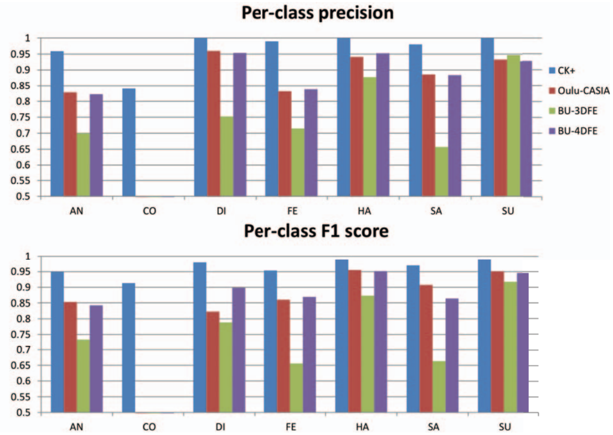

Figure 9. Per-class precision and F1 score on CK+, Oulu-CASIA, BU-3DFE and BU-4DFE databases

Per-expression precision and F1 score for CK+, Oulu-CASIA, BU-3DFE and BU-4DFE databases are shown in Fig.9. We observe that both precision and F1 score are the lowest in BU-3DFE database. This is due to the limited data size and high variety of ethnicities with larger range of ages, making it more challenging to recognize facial expressions.

V. CONCLUSIONS AND FUTURE WORKS

In this paper, we proposed an identity-adaptive training algorithm for facial expression recognition. First, six basic facial expression images are regenerated for any query image. Then, the facial expression is recognized by comparing the query image and the generated six facial images in the face space. Rather than trying to split image features into expression-related and identity-related parts, we keep the identity-related information adaptive to each query image, and make the facial expression recognition identity-independent.

For the future work, we plan to apply our method for facial expression recognition in the wild by training a more general facial expression regeneration model using more databases.

VI. ACKNOWLEDGMENTS

The material is based upon the work supported in part by the National Science Foundation under grants CNS-1629898 and CNS-1205664.

REFERENCES

[1] Z. Meng, P. Liu, J. Cai, S. Han, and Y. Tong, “Identity-aware convolutional neural network for facial expression recognition,” in *FG*. IEEE, 2017, pp. 558–565.

[2] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *CVPR*. IEEE, 2009, pp. 248–255.

[4] X. Zhao and X. e. a. Liang, "Peak-piloted deep network for facial expression recognition," in *ECCV*. Springer, 2016, pp. 425–442.

[5] H. Ding, S. K. Zhou, and R. Chellappa, "Facenet2expnet: Regularizing a deep face recognition net for expression recognition," in *Automatic Face & Gesture Recognition (FG 2017)*. IEEE, 2017, pp. 118–126.

[6] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *CVPR*, 2015, pp. 815–823.

[7] I. J. Goodfellow, D. Erhan, and P. L. e. a. Carrier, "Challenges in representation learning: A report on three machine learning contests," in *International Conference on Neural Information Processing*. Springer, 2013, pp. 117–124.

[8] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *Applications of Computer Vision (WACV)*. IEEE, 2016.

[9] J. Chen, X. Liu, P. Tu, and A. Aragones, "Learning person-specific models for facial expression and action unit recognition," *Pattern Recognition Letters*, vol. 34, no. 15, 2013.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

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

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

[13] J. Gauthier, "Conditional generative adversarial nets for convolutional face generation," *Class Project for Stanford CS231N: Convolutional Neural Networks for Visual Recognition, Winter semester*, vol. 2014, no. 5, p. 2, 2014.

[14] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *arXiv preprint arXiv:1611.07004*, 2016.

[15] P. Lucey and J. F. e. a. Cohn, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *CVPR Workshop*. IEEE, 2010.

[16] G. Zhao, X. Huang, and M. e. a. Taini, "Facial expression recognition from near-infrared videos," *Image and Vision Computing*, vol. 29, no. 9, pp. 607–619, 2011.

[17] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3d facial expression database for facial behavior research," in *FG*. IEEE, 2006, pp. 211–216.

[18] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale, "A high-resolution 3d dynamic facial expression database," in *FG*. IEEE, 2008, pp. 1–6.

[19] C. Ledig and L. e. a. Theis, "Photo-realistic single image super-resolution using a generative adversarial network," *arXiv preprint*, 2016.

[20] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *ECCV*. Springer, 2016, pp. 694–711.

[21] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *ECCV*. Springer, 2014, pp. 818–833.

[22] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015.

[23] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *CVPR*. IEEE, 2012, pp. 2879–2886.

[24] M. Abadi, A. Agarwal, and P. e. a. Barham, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.

[25] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 6, pp. 915–928, 2007.

[26] A. Klaser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *BMVC 2008-19th British Machine Vision Conference*. British Machine Vision Association, 2008, pp. 275–1.

[27] M. Liu, S. Li, S. Shan, R. Wang, and X. Chen, "Deeply learning deformable facial action parts model for dynamic expression analysis," in *Asian Conference on Computer Vision*. Springer, 2014, pp. 143–157.

[28] M. Liu, S. Shan, R. Wang, and X. Chen, "Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition," in *CVPR*, 2014, pp. 1749–1756.

[29] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *ICCV*, 2015, pp. 2983–2991.

[30] Y. Guo, G. Zhao, and M. Pietikäinen, "Dynamic facial expression recognition using longitudinal facial expression atlases," in *ECCV*. Springer, 2012, pp. 631–644.

[31] J. Wang, L. Yin, X. Wei, and Y. Sun, "3d facial expression recognition based on primitive surface feature distribution," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2. IEEE, 2006, pp. 1399–1406.

[32] S. Berretti, A. Del Bimbo, P. Pala, B. B. Amor, and M. Daoudi, "A set of selected sift features for 3d facial expression recognition," in *Pattern Recognition (ICPR)*. IEEE, 2010, pp. 4125–4128.

[33] X. Yang, D. Huang, Y. Wang, and L. Chen, "Automatic 3d facial expression recognition using geometric scattering representation," in *FG*, vol. 1. IEEE, 2015, pp. 1–6.

[34] A. T. Lopes and E. e. a. de Aguiar, "Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order," *Pattern Recognition*, vol. 61, pp. 610–628, 2017.

[35] A. Dapogny, K. Bailly, and S. Dubuisson, "Dynamic facial expression recognition by joint static and multi-time gap transition classification," in *FG*, vol. 1. IEEE, 2015.

[36] O. Rudovic, V. Pavlovic, and M. Pantic, "Multi-output laplacian dynamic ordinal regression for facial expression recognition and intensity estimation," in *CVPR*. IEEE, 2012, pp. 2634–2641.

[37] A. Dapogny, K. Bailly, and S. Dubuisson, "Pairwise conditional random forests for facial expression recognition," in *ICCV*, 2015, pp. 3783–3791.

[38] L. A. Jeni, D. Takacs, and A. Lorincz, "High quality facial expression recognition in video streams using shape related information only," in *ICCV Workshops*. IEEE, 2011, pp. 2168–2174.

[39] Z. Pan, M. Polceanu, and C. Lisetti, "On constrained local model feature normalization for facial expression recognition," in *International Conference on Intelligent Virtual Agents*, 2016, pp. 369–372.