14 Dec 2018

Facial Expression Recognition By De-expression Residue Learning

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

Umur Ciftci

Lijun Yin

Recommended Citation

H. Yang et al., "Facial Expression Recognition By De-expression Residue Learning," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2168 - 2177, article no. 8578329, Institute of Electrical and Electronics Engineers, Dec 2018.
The definitive version is available at https://doi.org/10.1109/CVPR.2018.00231

Facial Expression Recognition by De-expression Residue Learning

Huiyuan Yang, Umur Ciftci and Lijun Yin
Department of Computer Science
State University of New York at Binghamton, USA
{hyang51, uciftci}@binghamton.edu; lijun@cs.binghamton.edu

Abstract

A facial expression is a combination of an expressive component and a neutral component of a person. In this paper, we propose to recognize facial expressions by extracting information of the expressive component through a de-expression learning procedure, called De-expression Residue Learning (DeRL). First, a generative model is trained by cGAN. This model generates the corresponding neutral face image for any input face image . We call this procedure de-expression because the expressive information is filtered out by the generative model; however, the expressive information is still recorded in the intermediate layers. Given the neutral face image, unlike previous works using pixel-level or feature-level difference for facial expression classification, our new method learns the deposition (or residue) that remains in the intermediate layers of the generative model. Such a residue is essential as it contains the expressive component deposited in the generative model from any input facial expression images. Seven public facial expression databases are employed in our experiments. With two databases (BU-4DFE and BP4D-spontaneous) for pre-training, the DeRL method has been evaluated on five databases, CK+, Oulu-CASIA, MMI, BU-3DFE, and BP4D+. The experimental results demonstrate the superior performance of the proposed method.

1. Introduction

Research on facial expression recognition (FER) has been conducted on both posed and spontaneous facial expressions under various imaging conditions, including various head poses, illumination conditions, resolutions, and occlusions [34] [14] [17] [18] [12]. Although significant progress has been made towards improving the expression classification, the current main challenge comes from the large variations of individuals in attributes such as: age, gender, ethnic background and personality. Facial expressions may appear different (w.r.t. expressiveness or subtlety, etc.) for people with different personalities and expressive styles. Only recently have works [21] [13] started to take aspects of subject's identity attributes *i.e.*, age, gender, and personal characteristics, into consideration for facial expression analysis.

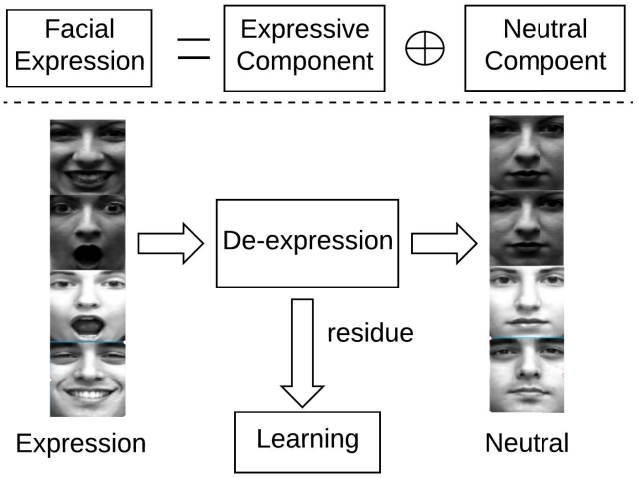

Figure 1. Illustration of our proposed method - De-expression Residue Learning (DeRL). A facial expression is the combination of a neutral face image and the expressive component. Our proposed method recognizes facial expression by learning the residual expressive component in the generative model.

Research shows that people are capable of recognizing facial expressions by comparing a subject's expression with a reference expression (i.e., neutral expression) of the same subject [4] [5] [10]. In other words, a facial expression can be decomposed to an expressive component and neutral component [25]. Up until now, several existing works utilized the image-difference or feature-difference of the query image and neutral face image [2] [30] [15] [13] to recognize facial expressions. However, their assumption is that the neutral expression must be obtainable. As a matter of fact, the neutral expression may not be always available for a given subject. In order to alleviate the problem, it is on demand to develop a neutral expression generator based on the given expressive input. The Generative Adversarial Model

2575-7075/18 $31.00 © 2018 IEEE
DOI 10.1109/CVPR.2018.00231

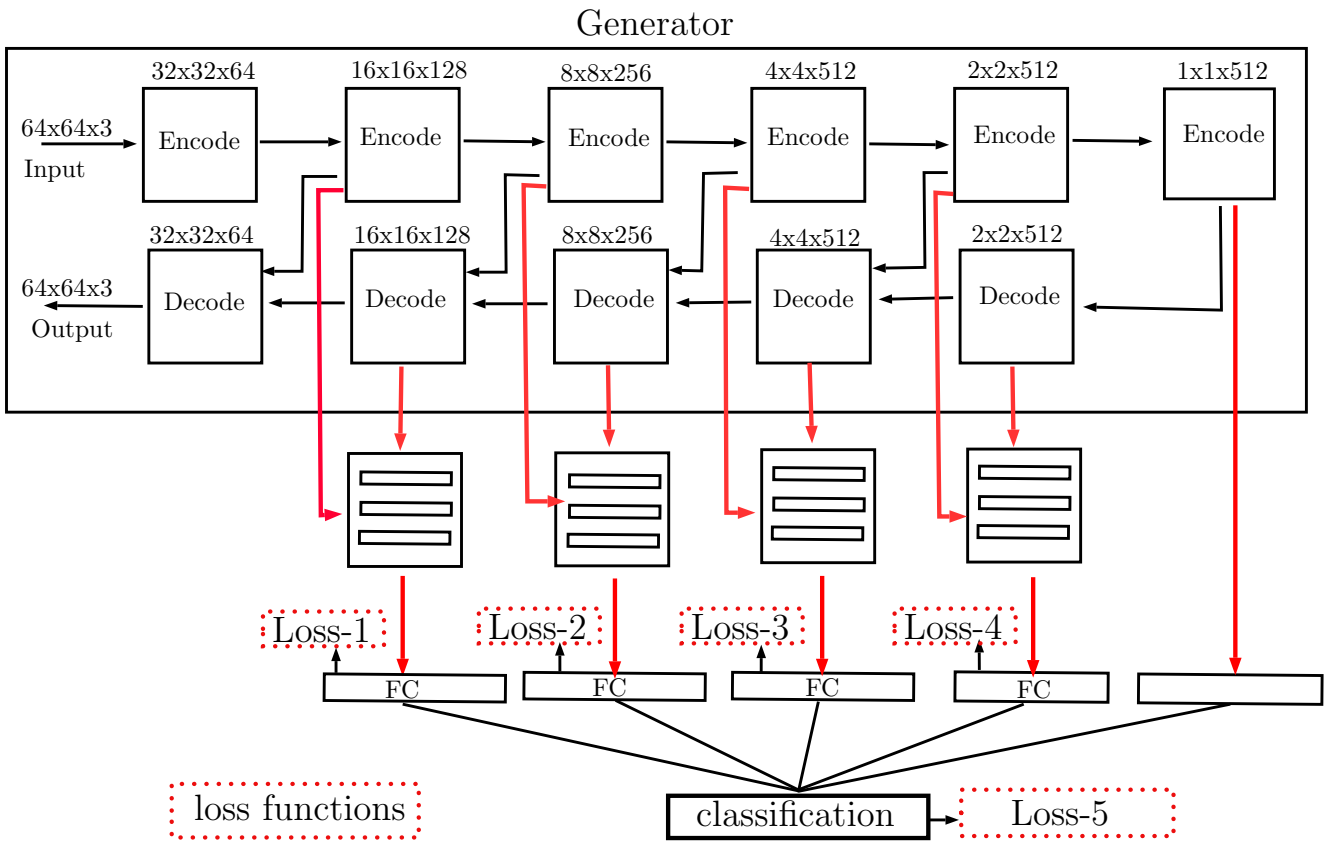

Figure 2. Framework of our proposed De-expression Residue Learning (DeRL) method.

(GAN) [8] is able to serve this purpose. To train a generative model (*generator*), a GAN framework utilizes another deep model (*discriminator*) to play an adversarial game with the generative model, rather than defining a regular cost function for the generator. The discriminator is designed to distinguish between samples from the generator and samples from the training data; while the generator learns to output samples that can maximally confuse the discriminator. An extension of the basic GAN is the *conditional* Generative Adversarial Networks (cGAN) [22], which is capable of learning different contextual information through the extra conditional variations. There are existing work that combine CNN and cGANs for many applications, including face generation [7], object reconstruction from edge maps [11], and object attributes manipulation [24].

In this paper, we propose a new approach called *De-exprssion Residue learning (DeRL)* to learn facial expressions by extracting the expressive component through a de-expression procedure. As illustrated in Fig.1, given a facial image with arbitrary expressions, its corresponding neutral expression is generated by the trained generative model. Through the procedure, the identity information of a subject remains unchanged while the expressive component is removed. We called this *de-expression*. Although the input image with an expression is "normalized" to the neutral expression as output, the expressive component that has been filtered out is still "deposited" in the generative model. In other words, the expression information is recorded in the generator during the de-expression procedure. Such a deposition is the residue of the de-expression, which is exactly the expressive component that we target to exploit for expression classification.

In contrast to the previous methods [15] [2] [30] [13], which used pixel-level difference or feature-level difference of expression images and neutral images, our proposed DeRL framework learns the de-expression residue remained in the generative model, with an attempt to mitigate the influence of individual identity and improve the performance of facial expression recognition. The contribution of this work lies in two-fold:

1. We propose a novel method to learn expressions by de-expression. We first train a generative model to generate the corresponding neutral face image for the query image; and then we learn the residue (*i.e.,* expressive component) of the generative model, thus alleviating the identity-related variation issue.

2. Our proposed method is capable of handling cases of both spontaneous expressions and posed expressions, with various styles and ethnic backgrounds. It successfully improves performance on individual datasets, as well as for cross-dataset validation, with better results compared to the state of the art.

2. Related Work

Previous works suggest that facial expression recognition could benefit from using a neutral face image[30] [13]. Subtracting a neutral face image from the corresponding facial expression image in either pixel-level or feature-level can emphasize the facial expression while reducing the intra-class variation.

Bazzo et al.[2] achieved good recognition rate by applying Gabor wavelets to recognize facial expression images subtracted from an averaged neutral face. Zafeiriou et al.[30] applied sparse facial expression representations for the difference images, which were derived from the subtraction of the neutral image from the expressive one, and demonstrated that the use of neutral images tends to emphasize the moved facial parts. Lee at al.[15] generated several intra-class variation images (including neutral) from a training set, which were then subtracted by the query face image to obtain difference images. The difference images were used to emphasize the facial expressions in a query face image. Kim et al.[13] employed a contrastive representation in the networks to extract the feature level difference between a query face image and a neutral face image.

However, these previous works made the assumption that the neutral expression is always available given any expression of the same subject, which is not realistic. It is on demand to generate a neutral face from any expression input. The recent utility of GAN shows success in such an application. Gauthier [7] tried to use cGANs to generate faces with specific attributes. Radford et al. [24] attempted to scale up GANs using CNNs to model images and introduced the structure of deep convolutional generative adversarial networks (DCGANs). This work showed the capability to manipulate the generated face samples by vector arithmetic. Isola et al. [11] utilized conditional adversarial networks for image-to-image translation and showed many interesting applications, *i.e.*, generating aerial photograph from map, reconstructing objects from edge maps, and colorizing images. Also, Zhou et al.[36] applied cGANs to synthesize facial expression images from the neutral faces.

So far, there has been use of image or feature difference of query images and the generated neutral images, but there is no exploration for any implicit expressive information recorded in the generative model. We propose to explore the expressive information, which is embedded in the generator, and extract the expression component from the intermediate layers directly. In fact, such information is "filtered out" by the generator during the de-expression process while its representation (or residue) is still deposited in the generative model, thus becoming the key information to represent the expressive component. Rather than using both query image and generated neutral face image to train a deep model with a contrastive loss function (e.g., [13]), our proposed method focuses on learning the residue of the generative model, leading to effectively capturing the expressive component and being more robust to individual variations.

3. Proposed Method - DeRL

The architecture of our proposed method - De-expression Residue Learning (DeRL), illustrated in Fig. 2, contains two learning processes: the first is learning the neutral face generation by cGANs, and the second is learning from the intermediate layers of the generator. The input image pairs, *e.g.* $< I_{input}, I_{target} >$, are used to train the cGANs. I_{input} is a face image showing any expression, and I_{target} is a neutral face image of the same subject. After training, the generator reconstructs the corresponding neutral face image for any input while keeping the identity information unchanged. From an expressive facial image to a neutral face image, the expression-related information is recorded as expressive component in the intermediate layers. For the second learning process, the parameters of the generator are fixed, and the output of the intermediate layers are combined and input into deep models for facial expression classification.

3.1. Neutral Face Regeneration

cGAN[22] is exploited to generate a neutral face representation from a given expressive face image . A GAN framework usually contains two different players: a generator (G) and discriminator (D). The generator is trained to recover the distribution of the training data by playing a so-called minmax game with the discriminator. Image pairs $< I_{input}, I_{target} >$ are provided for training the cGANs. I_{input} is first input into the generator to reconstruct I_{output}, and then $< I_{input}, I_{target}, yes >$ and $< I_{input}, I_{output}, no >$ are given to the discriminator. The discriminator tries to distinguish the $< I_{input}, I_{target} >$ from the $< I_{input}, I_{output} >$, while the generator tries to not only maximally confuse the discriminator but also generate an image as close to the target image as possible.

The objective for the discriminator is expressed as:

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

, where *N* is the total number of training image pairs. The

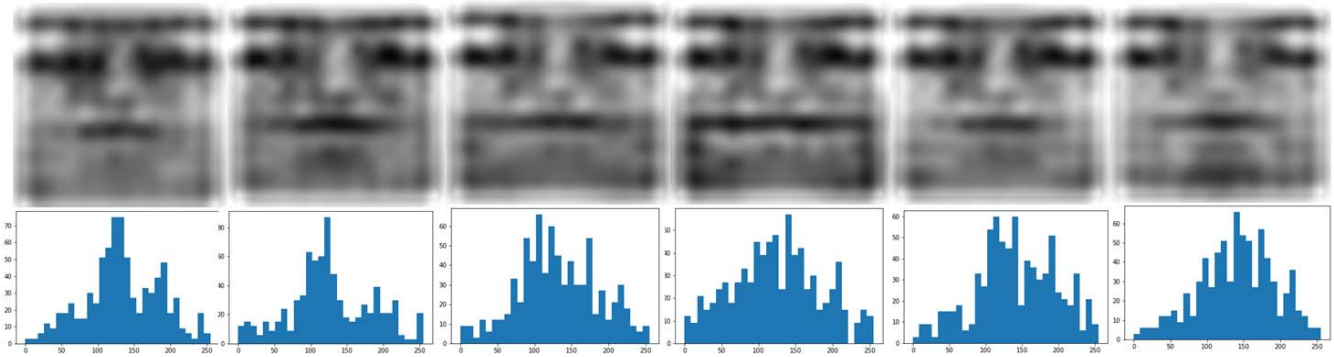

Figure 3. Illustration of De-expression Residue, which are the expressive components for *anger, disgust, fear, happiness, sadness, and surprise*, from left and right, respectively. The corresponding expressive component histogram is also shown in the second row.

objective for the generator is described below:

$$L_{cGAN}(G) = -\frac{1}{N}\sum_{i=1}^{N}\Big\{log\Big(D\big(I_{input}, G(I_{input})\big)\Big) + \theta_1 \cdot ||I_{target} - G(I_{input})||_1\Big\} \quad (2)$$

Here, we use L1 loss for the image similarity rather than L2, because L2 loss is prone to over-blurring the output image [11]. The final objective is:

$$G^* = arg\min_G \max_D L_{cGAN}(D) + \theta_2 \cdot L_{cGAN}(G) \quad (3)$$

3.2. Learning Facial Expressive Component

After the neutral face regeneration, the expression information can be analyzed by comparing the neutral face and the query expression face at pixel level or feature level. However, the pixel level difference is unreliable due to the variation between images *i.e.*, rotation, translation and lighting condition changes. This can cause a large pixel-level difference even without expression changes. Also, the feature level difference is unstable, as the expression information may vary according to the identity information. Since the *difference* of the query image and the neutral image is recorded in the intermediate layers, we exploit the expressive component from the intermediate layers directly to alleviate the above problem, .

We denote an image with a facial expression as: I^{id}_{exp}. After it is input into the generative model, a neutral expression image is generated:

$$I^{id=A}_{exp=neutral} = G\Big(I^{id=A}_{exp=E}\Big) \quad (4)$$

where,G is the generator, E belongs to any of the six basic prototypic facial expressions. From the Equation (4), we can see that an image with subject (A) and expression (E) becomes a neutral face of the same subject (A). It is reasonable to conclude that the unique expression information (a.k.a. expressive component) of each individual must be recorded in the intermediate layers of the generator. Therefore, we propose the second learning strategy, which is to learn expressions from the intermediate layers of the generator directly. This unique information is also named as *De-Expression Residue* (see Fig. 3 for examples).

As shown in Fig. 2, in order to learn the de-expression residue from the intermediate layers of the generator, all the filters of those layers are fixed, and all the layers that have the same size are concatenated and input into a local CNN model for facial expression classification. For each local CNN model, the cost function is noted as $loss_i, i \in [1, 2, 3, 4]$. The last fully connected layers of each local CNN model are further concatenated and combined with the last encode layer for facial expression classification. Consequently, the total loss function is defined as:

$$total\_loss = \lambda_1 loss_1 + \lambda_2 loss_2 + \lambda_3 loss_3 + \lambda_4 loss_4 + \lambda_5 loss_5 \quad (5)$$

4. Experiments

The proposed DeRL approach is evaluated on five public facial expression databases, including CK+ [20], Oulu-CASIA [33], MMI [23] and BU-3DFE [29], and spontaneous expression database BP4D+ [32]. Additionally, two other databases with posed expressions BU-4DFE [28] and spontaneous expressions BP4D [31] are used for pre-training.

4.1. Implementation Details

Three landmark points (the centers of eyes, and the chin) are used to align the face region. For the databases where landmarks are not provided, we apply the TSM [37] for face detection and landmark localization. The aligned face region is then cropped and resized to the size of 70×70. To avoid over-fitting, we apply a data augmentation method to generate more training data. First, five 64×64 patches are

cropped-out from five locations of each image (*center and four corners, respectively*). Each image patch is rotated by $[-15^o, -12^o, -9^o, -6^o, -3^o, 0^o, 3^o, 6^o, 9^o, 12^o, 15^o]$ respectively. Horizontally flipping is also applied. The result is an augmented dataset, which is 110 times larger than the original one. The data augmentation method is only applied to the training data.

The generative model is first pre-trained on the BU-4DFE [28], which contains 60,600 images from 101 subjects. Each subject has six sequences, and each sequence shows one of the six basic facial expressions from neutral to peak expression, ending with neutral again. To construct the training dataset, the first frame of each sequence is used as target image, and the rest of the sequence is used as input images. The pre-trained model is further fine-tuned on CK+, Oulu-CASIA, MMI, BU-3DFE, and BP4D+ databases.

We use the Adam optimizer with a batch size of 150, momentum of 0.9, and dropout of 0.5 for the fully connected layers during training. We use 200 epochs to train the generative models, and 50 epochs for the facial expression classification models. We set $\lambda_1 = 0.7$, $\lambda_2 = 0.5$, $\lambda_3 = 0.3$, $\lambda_4 = 0.2$ and $\lambda_5 = 1.0$ for the loss function respectively. The proposed method is implemented using *tensorflow* [1] on the GeForce GTX 1080 platform.

4.2. Visualization of Expressive Component and Regenerated Neutral Faces

Fig.4 illustrates several samples of the generated neutral face image on CK+, Oulu-CASIA, BU-4DFE, and BP4D+ databases, respectively. The first column is the input image, the third column is the ground-truth neutral face image of the same subject, and the middle is the output of the generative model. As shown, the expressive component is filtered out by the generative model while the subject-related information is preserved.

Fig.3 illustrates the samples of the de-expression residue from the CK+ dataset, which are the expressive components for *anger, disgust, fear, happiness, sadness, and surprise*, respectively. The corresponding histogram of each expressive component is also displayed. As we can see, both expressive components and corresponding histograms are distinguishable among the six expressions.

4.3. Expression Recognition Results

The **Extended Cohn-Kanade database (CK+)** [20] is widely used for evaluating facial expression recognition. It contains 593 video sequences collected from 123 subjects. Among them, 327 video sequences with 118 subjects are labeled as one of seven expressions, *i.e.* anger, contempt, disgust, fear, happiness, sadness and surprise, from neutral to peak expression. Only the last frame of each sequence is labeled, as a general procedure, we use the last three frames

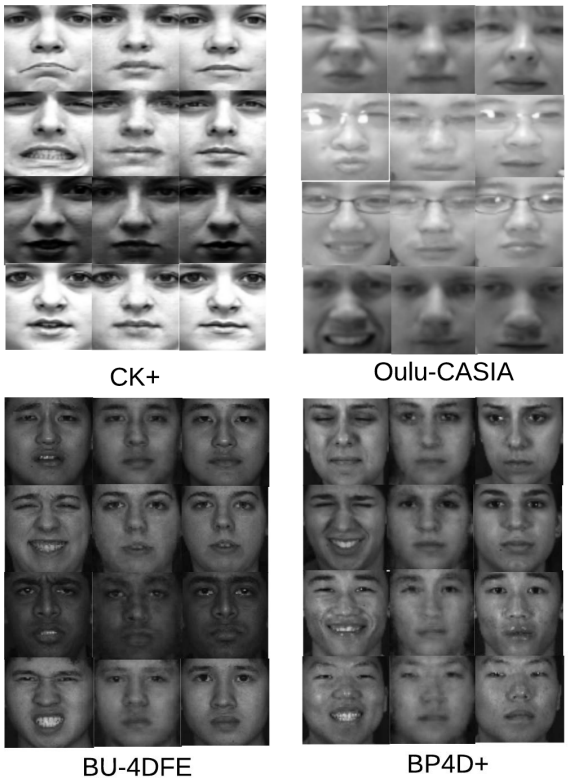

Figure 4. Illustration of neutral faces generated by the generative model for the input query images from CK+, Oulu-CASIA, BU-4DFE and BP4D+ databases, respectively.

of each sequence with the provided label, which results in 981 images. The images are further split into 10 folds based on the identity in an ascending order, thus the subjects in any two subsets are mutually exclusive.

The results are reported as the average of the 10 runs. As shown in Table 1, our proposed DeRL method achieves over 97% recognition rate, outperforming the compared state-of-the-art methods. Note that all these methods except the IACNN [21], CNN baseline and our DeRL method exploited temporal information extracted from image sequences. DTAGN [12] also utilized the landmark features. In contrast to all other methods, our proposed DeRL method performs well on recognizing facial expressions on static images and achieves around 2% improvement compared to IACNN [21]. Fig. 5 is the confusion matrix of our method, where *fear* expression shows the lowest recognition rate with 90%.

The **Oulu-CASIA database** [33] contains data captured under three different illumination conditions using two types of cameras. During the experiment, only the data captured under strong illumination condition with the VIS camera is used. The Oulu-CASIA VIS has 480 video sequences taken from 80 subjects, and each video sequence

Table 1. Average accuracy on the CK+ database for seven expressions classification.

Method	Setting	Accuracy
LBP-TOP [34]	sequence-based	88.99
HOG 3D [14]	sequence-based	91.44
3DCNN [17]	sequence-based	85.9
STM-Explet [18]	sequence-based	94.19
IACNN [21]	image-based	95.37
DTAGN [12]	sequence-based	97.25
CNN (baseline)	image-based	89.50
DeRF (Ours)	image-based	**97.30**

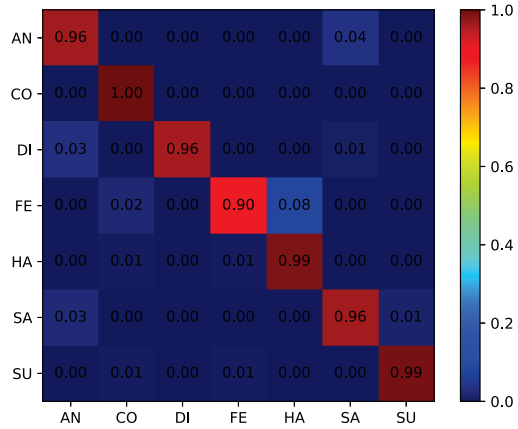

Figure 5. Confusion matrix on CK+

Table 2. Average accuracy on the Oulu-CASIA database for six expressions classification.

Method	Setting	Accuracy
LBP-TOP [34]	sequence-based	68.13
HOG 3D [14]	sequence-based	70.63
STM-Explet [18]	sequence-based	74.59
Atlases [9]	sequence-based	75.52
DTAGN-Joint [12]	sequence-based	81.46
FN2EN [6]	image-based	87.71
PPDN [35]	image-based	84.59
CNN (baseline)	image-based	72.92
DeRL (Ours)	image-based	**88.0**

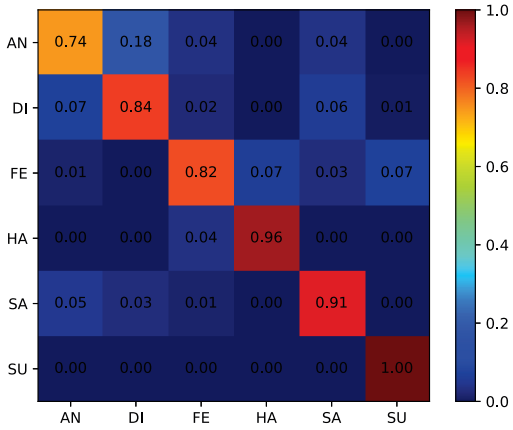

Figure 6. Confusion matrix on Oulu-CASIA

is labeled as one of the six basic expressions. Similar to the CK+ database, each video sequence starts from a neutral face and ends with a peak facial expression. To include more data, the last three frames of each sequence are selected. Similar to the experimental setting in CK+, a 10-fold subject-independent cross validation is performed.

The average accuracy on recognizing six expressions on Oulu-CASIA over 10 runs is shown in Table.2. Our DeRL method achieves the highest accuracy compared to those of state-of-the-art methods, including CNN-based methods (DTAGN-Joint [12], FN2EN [6], and PPDN [35]) and handcrafted features based methods (LBP-TOP [34], HOG 3D [14] and STM-Explet [18]). Note that only FN2EN [6], PPDN [35] and our DeRL method use the static image for facial expression recognition, while others exploited the temporal information of video sequences. The confusion matrix is shown in Fig. 6, our method performs very well in recognizing expressions *happiness* and *surprise*, while *anger* shows the relatively low recognition rate, which is mostly confused with *disgust*.

The **MMI database** [23] consists of 236 image sequences from 31 subjects. Each sequence is labeled as one of the six basic facial expressions. We selected 208 sequences captured in frontal view. Each sequence starts with a neutral expression, reaches peak expression near the middle of the sequence, and ends with a neutral expression. Since the label is only given for the whole sequence, we selected three frames in the middle of each sequence as peak frames and associated them with the provided labels. This results a dataset with 624 images. Similar to the CK+ database setting, a 10-fold subject-independent cross validation is performed.

Table. 3 reports the average accuracy of 10 runs on the MMI database for recognizing six expressions. Although STM-Explet [18] shows the highest accuracy of 75.12%, it employs the temporal information extracted from the video sequence. Our DeRL method, showing a close result of 73.23%, recognizes facial expressions based only on static images, which is more suitable for some applications where video sequences are not available. Compared to the IACNN [21], which is also an image-based method, our DeRL

Table 3. Average accuracy on the MMI database for six expressions classification.

Method	Setting	Accuracy
LBP-TOP [34]	sequence-based	59.51
HOG 3D [14]	sequence-based	60.89
STM-Explet [18]	sequence-based	**75.12**
DTAGN-Joint [12]	sequence-based	70.24
IACNN [21]	image-based	71.55
CNN (baseline)	image-based	57.00
DeRL (Ours)	image-based	**73.23**

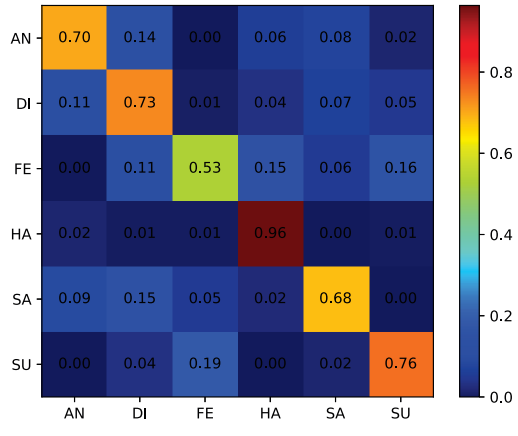

Figure 7. Confusion matrix on MMI

Table 4. Average accuracy on the BU-3DFE database for six expressions classification.

Method	Setting	Accuracy
Wang et al.[26]	3D	61.79
Berretti et al.[3]	3D	77.54
Yang et al.[27]	3D	84.80
Lo et al.[16]	2D image + 3D	**86.32**
Lopes [19]	image-based	72.89
CNN (baseline)	image-based	73.2
DeRL (Ours)	image-based	**84.17**

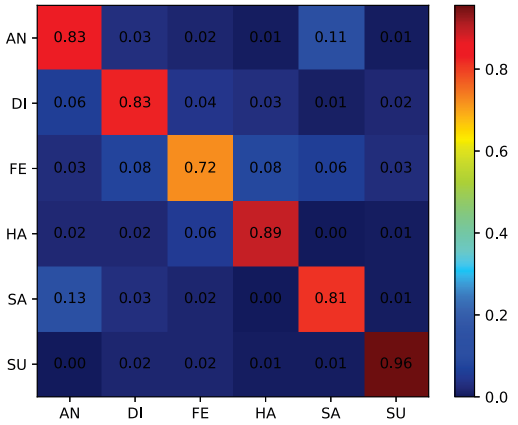

Figure 8. Confusion matrix on BU-3DFE

method shows improvement of 1.7%. As shown from the confusion matrix in Fig. 7, *fear* is relatively hard for recognition, mostly confused with *surprise* and *disgust*. On the other hand, *happiness* is relatively easy to classify.

The **BU-3DFE database** [29] contains 2,500 pairs of static 3D face models and texture images from 100 subjects with a variety of ages and races. For each subject, six basic expressions with four levels of intensity and a neutral expression, was captured. During the experiment, only the texture images and the high intensity expressions (i.e. the last two levels) are used. A 10-fold cross-validation is performed, and the split is subject independent.

Table 4 summarizes the average result of 10 runs on the BU-3DFE database. From the data we can find that our proposed method outperforms the other two image-based methods significantly. Notice that the multi-modality based approach (2D+3D) [16] performs better than the single modality approaches. However, our single modality (image-based) DeRL method achieves close performance compared to the multi-modal fusion approach. As we can see from the confusion matrix Fig. 8 on BU-3DFE database, *surprise* is relatively easy to recognize, showing a 96% recognition rate, while *fear* has a relatively low recognition rate, and mostly confused with *disgust* and *happiness*.

The **BP4D+** [32] is a multimodal spontaneous emotion corpus (MMSE), including synchronized 3D, 2D, thermal, physiological data sequences (e.g., heart rate, blood pressure, skin conductance (EDA), and respiration rate) from 140 subjects (58 males and 82 females) with ages ranging from 18 to 66 years old. Although the database provides FACS codes, there is no facial expression label available for each frame. In order to evaluate the spontaneous expression database BP4D+, we semi-automatically select 2468 frames from 72 subjects (45 female and 27 male) on four tasks based on the FACS codes. In the experiment, we only use the 2D texture images and four kind of expressions (*i.e., happiness, surprise, pain, and neutral*). A 10-fold cross-validation is performed, and the split is subject independent. As we can see in Table.5, our proposed method outperforms the CNN baseline when both training and testing are done on BP4D+.

To further validate our DeRL approach, we have also conducted a cross-database validation on recognizing four expressions (*i*.e., happiness, surprise, pain, and neutral). We choose the spontaneous expression database BP4D [31] for

Table 5. Average accuracy on the BP4D+ database for four expressions classification.

Method	Setting	Accuracy
CNN (baseline)	image-based	76.5
DeRL (ours) train on BP4D, test on BP4D+	image-based	**74.41**
DeRL (ours) train and test on BP4D+	image-based	**81.39**

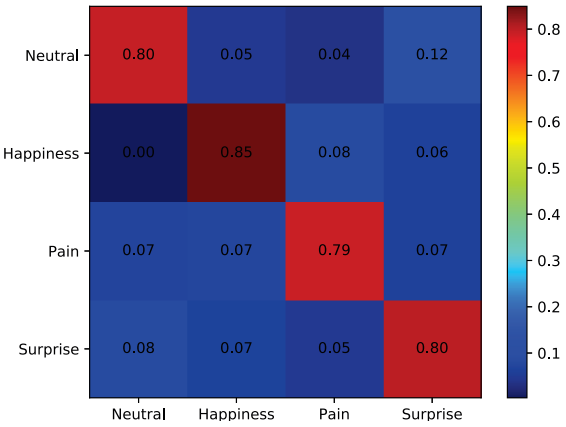

Figure 9. Confusion matrix of recognizing four expressions on BP4D+

training, and BP4D+ for testing. Similar to the labeling process on BP4D+, we also labeled 1262 frames from all 41 subjects in the BP4D database based on the AU codes in a semi-automatic way. The experiment result shows that the performance on cross-database validation is lower than the same-database validation, nevertheless, it is still comparable with the CNN baseline.

4.4. Classification from Internal Layers

Internal layers of the generative model have different contributions to the recognition rate. As we can see in Fig. 10, the local CNN-1 and CNN-2 models show much higher recognition rates than the CNN-3 and CNN-4 . Thus it justifies the λ_1 and λ_2 have bigger weights than λ_3 and λ_4 for combination of total loss in Equation (5). Such a combination achieves the highest recognition rate on the individual dataset.

5. Conclusion

In this paper, we present a novel approach for facial expression recognition, which is based on the de-expression residue learning (a.k.a. DeRL). First, a generative model is trained by cGANs to regenerate the neutral face image for any query image. Second, a learning procedure is performed on the internal layers of the generative model. The learning procedure is able to capture the expressive component of facial expressions that have been recorded in the generative model.

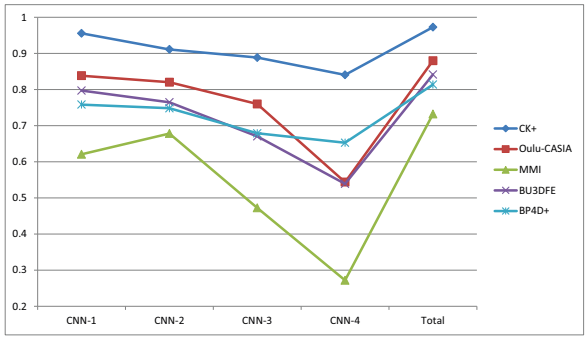

Figure 10. Recognition rates based on individual input parts and their combination.

Our proposed method was evaluated on both posed and spontaneous expressions datasets. Without exploiting the temporal information, the DeRL method outperforms the baseline CNN methods, the state-of-the-art image-based methods, and even most of the state-of-the-art sequence-based methods that utilize the spatio-temporal information. A cross-database validation is also performed to show the effectiveness of extracting the expressive component from the intermediate layers of the generative model. Such expressive components are extendible for facial action units detection. Our future work will incorporate the expressive components with temporal information for addressing the issues of AU detection, as well as 3D face analysis with a variety of head poses and subtle facial changes.

6. Acknowledgement

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

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

[2] J. J. Bazzo and M. V. Lamar. Recognizing facial actions using gabor wavelets with neutral face average difference. In *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*, pages 505–510. IEEE, 2004.

[3] 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), 2010 20th International Conference on*, pages 4125–4128. IEEE, 2010.

[4] V. Bruce and A. Young. Understanding face recognition. *British journal of psychology*, 77(3):305–327, 1986.

[5] A. J. Calder and A. W. Young. Understanding the recognition of facial identity and facial expression. *Nature Reviews Neuroscience*, 6(8):641–651, 2005.

[6] 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), 2017 12th IEEE International Conference on*, pages 118–126. IEEE, 2017.

[7] J. Gauthier. Conditional generative adversarial nets for convolutional face generation. *Class Project for Stanford CS231N: Convolutional Neural Networks for Visual Recognition, Winter semester*, 2014(5):2, 2014.

[8] 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*, pages 2672–2680, 2014.

[9] Y. Guo, G. Zhao, and M. Pietikäinen. Dynamic facial expression recognition using longitudinal facial expression atlases. In *Computer Vision–ECCV 2012*, pages 631–644. Springer, 2012.

[10] J. V. Haxby, E. A. Hoffman, and M. I. Gobbini. The distributed human neural system for face perception. *Trends in cognitive sciences*, 4(6):223–233, 2000.

[11] 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.

[12] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim. Joint fine-tuning in deep neural networks for facial expression recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2983–2991, 2015.

[13] Y. Kim, B. Yoo, Y. Kwak, C. Choi, and J. Kim. Deep generative-contrastive networks for facial expression recognition. *arXiv preprint arXiv:1703.07140*, 2017.

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

[15] S. H. Lee, K. N. K. Plataniotis, and Y. M. Ro. Intra-class variation reduction using training expression images for sparse representation based facial expression recognition. *IEEE Transactions on Affective Computing*, 5(3):340–351, 2014.

[16] H. Li, H. Ding, D. Huang, Y. Wang, X. Zhao, J.-M. Morvan, and L. Chen. An efficient multimodal 2d+ 3d feature-based approach to automatic facial expression recognition. *Computer Vision and Image Understanding*, 140:83–92, 2015.

[17] 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*, pages 143–157. Springer, 2014.

[18] M. Liu, S. Shan, R. Wang, and X. Chen. Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1749–1756, 2014.

[19] A. T. Lopes, E. de Aguiar, A. F. De Souza, and T. Oliveira-Santos. Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order. *Pattern Recognition*, 61:610–628, 2017.

[20] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 94–101. IEEE, 2010.

[21] Z. Meng, P. Liu, J. Cai, S. Han, and Y. Tong. Identity-aware convolutional neural network for facial expression recognition. In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 558–565. IEEE, 2017.

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

[23] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 5–pp. IEEE, 2005.

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

[25] H. Wang et al. Facial expression decomposition. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 958–965. IEEE, 2003.

[26] 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*, volume 2, pages 1399–1406. IEEE, 2006.

[27] X. Yang, D. Huang, Y. Wang, and L. Chen. Automatic 3d facial expression recognition using geometric scattering representation. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 1, pages 1–6. IEEE, 2015.

[28] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale. A high-resolution 3d dynamic facial expression database. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, pages 1–6. IEEE, 2008.

[29] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato. A 3d facial expression database for facial behavior research. In *Automatic face and gesture recognition, 2006. FGR 2006. 7th international conference on*, pages 211–216. IEEE, 2006.

[30] S. Zafeiriou and M. Petrou. Sparse representations for facial expressions recognition via l 1 optimization. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 32–39. IEEE, 2010.

[31] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014.

[32] Z. Zhang, J. M. Girard, Y. Wu, X. Zhang, P. Liu, U. Ciftci, S. Canavan, M. Reale, A. Horowitz, H. Yang, et al. Multi-

modal spontaneous emotion corpus for human behavior analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3438–3446, 2016.

[33] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikäinen. Facial expression recognition from near-infrared videos. *Image and Vision Computing*, 29(9):607–619, 2011.

[34] 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*, 29(6):915–928, 2007.

[35] X. Zhao, X. Liang, L. Liu, T. Li, Y. Han, N. Vasconcelos, and S. Yan. Peak-piloted deep network for facial expression recognition. In *European Conference on Computer Vision*, pages 425–442. Springer, 2016.

[36] Y. Zhou and B. E. Shi. Photorealistic facial expression synthesis by the conditional difference adversarial autoencoder. *arXiv preprint arXiv:1708.09126*, 2017.

[37] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2879–2886. IEEE, 2012.