Facial motion prior networks for facial expression recognition

Chen, Yuedong; Wang, Jianfeng; Chen, Shikai; Shi, Zhongchao; Cai, Jianfei

2019

Generative Adversarial Networks (GANs) to generate neutral faces by learning and filtering the expressive information which is later used for facial expression classification.

Despite great progress, the existing deep learning based FER methods still have some limitations. Firstly, most methods do not incorporate domain knowledge well, so the global features they extract tend to be less discriminative and less representative for FER. Secondly, although some deep learning based methods consider domain knowledge to extract local geometry information by assuming the availability of the full set of landmarks, optical flows or AUs, their assumptions might not be valid since in many cases we might not have all the extra local geometry information and the AU detection task itself is a challenging one. Finally, GANs based method is also impractical since it requires the neutral and expressive faces of the same person are simultaneously available for all the training data. In addition, the image quality of the generated faces is hard to control, which directly affects the performance of the subsequent expression classifier.

Therefore, to address these problems, we propose a novel FER framework. Our contribution can be summarized as follows. Firstly, we design a novel end-to-end deep learning framework named Facial Motion Prior Networks (FMPN) for FER, where we introduce an addition stream to generate a mask to focus on facial muscle moving regions. Secondly, we incorporate prior domain knowledge by using the average differences between neutral faces and the corresponding expressive faces as the guidance for the facial motion mask learning. Finally, our method outperforms current state-of-the-art results on two laboratory-controlled datasets and one in-the-wild dataset, which demonstrates the effectiveness of the proposed framework.

II. PROPOSED APPROACH

An expressive face of a person is a deformation from the neutral face. In other words, the differences between a neutral face and its corresponding expressive face contain lots of information related to facial expressions. Since basic facial expressions, such as *anger*, *fear*, *happiness* and so on, share traits of uniformity across different people and races, it is reasonable to learn to emphasize specific facial moving parts when conducting expression recognition. On the other hand, emphasizing local facial parts (moving muscle) should not lead

Abstract—**Deep learning based facial expression recognition (FER) has received a lot of attention in the past few years. Most of the existing deep learning based FER methods do not consider domain knowledge well, which thereby fail to extract representative features. In this work, we propose a novel FER framework, named Facial Motion Prior Networks (FMPN). Particularly, we introduce an addition branch to generate a facial mask so as to focus on facial muscle moving regions. To guide the facial mask learning, we propose to incorporate prior domain knowledge by using the average differences between neutral faces and the corresponding expressive faces as the training guidance. Extensive experiments on three facial expression benchmark datasets demonstrate the effectiveness of the proposed method, compared with the state-of-the-art approaches.**

Index Terms—**facial expression recognition, deep learning, prior knowledge, facial-motion mask**

I. INTRODUCTION

Facial expression is one of the most important components in daily communications of human beings. It is generated by movements of facial muscles. While different people have different kinds of facial expressions caused by their own expressive styles or personalities, many studies show that there are several types of basic expressions shared by different peoples with different cultural and ethnic background [1]. Research on automatic recognition of such basic facial expressions has drawn great attention during the past decades.

Traditional approaches tend to conduct facial expression recognition (FER) by using Gabor Wavelets, sparse coding, etc., where many studies show that subtracting neutral faces from their corresponding expressive faces can help the algorithms to emphasize on the facial moving areas, and significantly improve the expression recognition rate [2].

Convolutional Neural Networks (CNNs) have been widely applied to FER in the recent years. CNNs can achieve good performance by learning powerful high-level features which are better than those conventional hand-crafted features. There are also some other methods proposing to combine global appearance features with local geometry features for FER. Specifically, they feed facial expression recognition networks with not only the original images, but also their related facial landmarks [3], or optical flow [4]. [5] extracts more precise facial features by focusing on some specific local parts, inspired by Action Units (AUs) [6], [7]. Recently, a de-expression framework was proposed in [8]. They used

Yuedong Chen[*], Jianfeng Wang[†], Shikai Chen[‡], Zhongchao Shi[†] and Jianfei Cai[§]
[*]Nanyang Technological University, Singapore, [†]AI Lab, Lenovo Research, China,
[‡]Southeast University, China, [§]Monash University, Australia
donald.chen@ntu.edu.sg, {wangjf17, shizc2}@lenovo.com, skchen@seu.edu.cn, jianfei.cai@monash.edu
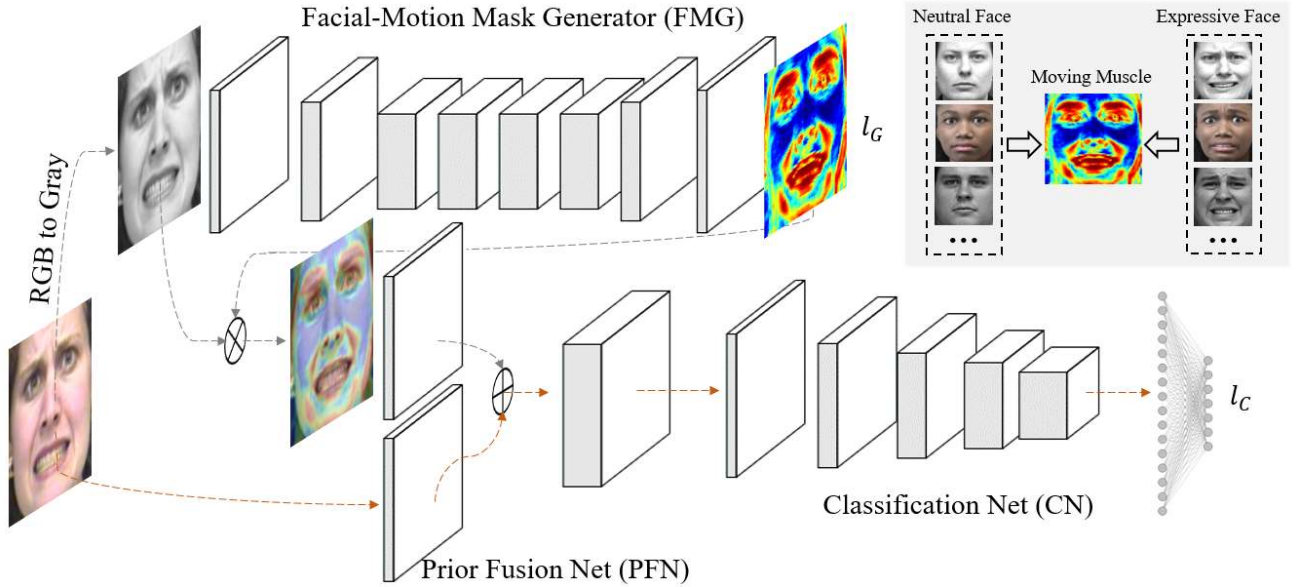
Fig. 1. Architecture of the proposed method. The model is composed of three networks: Facial-Motion Mask Generator (FMG), Prior Fusion Net (PFN) and Classification Net (CN). An expressive face is converted to gray scale and fed to FMG to generate a facial-motion mask. Then the mask is applied to and fused with the original input expressive face in PFN. The output of PFN is further fed to CN to extract more powerful features and predict facial expression label. l_G and l_C are loss functions at FMG and CN, respectively, which are end-to-end jointly optimized during training. Note that the learning of FMG is guided by pseudo ground truth masks, which are the average differences between neutral faces and their corresponding expressive faces (see top right corner).

to totally ignoring holistic facial image (whole face), because attributes such as gender and age, provided by the whole face image, can also affect types of expressions significantly. Therefore, both local features and holistic features should be taken into consideration. Inspired by the above analysis, we propose a novel approach to recognize facial expressions.

Fig. 1 shows the architecture of the proposed FMPN framework, which consists of three networks: **Facial-Motion Mask Generator (FMG)**, **Prior Fusion Net (PFN)** and **Classification Net (CN)**. FMG is constructed to generate a mask, namely facial-motion mask, which highlights moving areas of the given gray scale expressive face. PFN aims to fuse the original input image with the facial-motion mask generated by FMG to introduce domain knowledge to the whole framework. CN is a typical Convolutional Neural Network (CNN) for feature extraction and classification, such as VGG, ResNet or Inception. We mainly discuss FMG and PFN in the following subsections.

Facial-Motion Mask Generator (FMG) is built to generate a facial-motion mask, which is used to highlight expression-related facial motion regions. Instead of making the network to learn active areas blindly, we choose to guide it via some pseudo ground-truth masks that are generated by modeling basic facial expressions.

In particular, as aforementioned, facial expressions are caused by the contraction of facial muscles, and different people with the same expression share a similar pattern. Therefore, for one specific type of facial expressions, we model muscle moving areas as the difference between an expressive face and its corresponding neutral face, while the characteristic of similarity is modeled by averaging the above

differences of all the training instances in the same expression category. Specifically, for a k-th type of facial expressions, e.g., *happiness*, its ground truth mask $I_m^{(k)}$ is constructed as

$$I_m^{(k)} = \varphi\left(\frac{1}{N_k} \sum_i^{N_k} \left| \xi(R_{e,i}^{(k)}) - \xi(R_{n,i}^{(k)}) \right|\right), \quad (1)$$

where $R_e^{(k)}$ is the unprocessed/raw face with the k-th type of facial expressions, $R_n^{(k)}$ is the corresponding neutral face (faces of the same person shares one neutral face), N_k is the number of faces in the k-th expression category, and $\xi(\cdot)$ and $\varphi(\cdot)$ refer to the pre-processing and the post-processing, respectively.

Since facial moving areas are directly modeled as the absolute differences in Eq. (1), the expressive and neutral faces need to be well aligned, which is ensured by the pre-processing function $\xi(\cdot)$. Specifically, a standard spatial transformation is performed by aligning the detected facial landmarks with the standard reference landmarks. In addition, considering that the average face difference is often with close-contrast values, we further introduce a post-processing function $\varphi(\cdot)$ which applies histogram equalization to adjust the difference values for a better distribution. The generated ground truth masks of CK+ [9] for the seven basic expressions are shown in Fig. 2.

Given an expressive face, FMG is designed to learn a facial-motion mask, trained with the guidance from the pre-computed ground truth mask related to that expression. Particularly, as illustrated in Fig. 1, the input expressive face is first down-sampled and convoluted to extract feature related to geometric structure through two convolutional layers. Then the geometric feature is filtered and transferred to semantic feature related

Fig. 2. Ground truth masks of the CK+. Corresponding expressions from left to right, top to bottom are: *anger*, *contempt*, *disgust*, *fear*, *happiness*, *sadness* and *surprise*. It can be seen that different facial expressions have different moving muscles, which can benefit the recognition of expressions.

to the dynamic area through four residual blocks. Finally, two transposed convolutional layers are appended to project the learned semantic feature back to spatial domain as moving muscle focused mask. We use MSE (mean square error) for the training objective function of FMG:

$$l_G(I_e, k) = \mathbb{E}(f_G(I_e) - I_m^{(k)})^2, \qquad (2)$$

where I_e is the input expressive face, $I_m^{(k)}$ is the ground truth mask corresponding to the expression class of I_e, defined in (1), and $f_G(I_e)$ refers to the mask generated by FMG.

One might ask why we do not directly use the computed ground truth masks for facial expression recognition, instead of learning to generate a facial-motion mask. One main reason is that during testing we do not know which ground truth mask should be chosen since different expressions have different facial-motion masks. We would also like to point out that, considering expressive faces with the same expression in different datasets have similar moving muscles, the ground truth masks obtained from one dataset are mostly likely to be suitable for another dataset. This can help overcome the limitation that some dataset may not contain paired expressive and neutral faces to compute ground truth masks.

Prior Fusion Net (PFN) is designed to automatically fuse the original input face with the face that is masked by the facial-motion mask learned from FMG. The former is to provide holistic features, while the latter emphasizes the moving areas, which reflects the common expression definition and domain knowledge. Specifically, PFN produces a fused output I_s by a weighted sum, which can be written as

$$I_s = w_1 \cdot I_{e'} + w_2 \cdot (I_e \otimes f_G(I_e)), \qquad (3)$$

where $I_{e'}$ is the RGB version of the gray-scale image I_e, $I_e \otimes f_G(I_e)$ refers to the masked face, which is obtained by element-wise multiplication of face I_e and its corresponding mask $f_G(I_e)$, and w_1 and w_2 are weights of convolutional layers, which are updated during training.

After PFN, the fused output I_s will then be fed into a CNN based classification network, which can be VGG, ResNet or

others. The classification network is trained with the cross entropy loss:

$$l_C(I_s, k) = -\log\left(\frac{\exp(f_C(I_s)^{(k)})}{\sum_i^K \exp(f_C(I_s)^{(i)})}\right), \qquad (4)$$

where $f_C(I_s)^{(i)}$ is the i-th output value of the classification network $f_C(\cdot)$, k is the target expression, and K is the total number of facial expression classes in a given dataset. In this way, the total loss becomes

$$l_{total} = \lambda_1 \cdot l_G + \lambda_2 \cdot l_C, \qquad (5)$$

where l_G and l_C are defined in (2) and (4), respectively, and λ_1 and λ_2 are hyperparameters, being empirically set in training.

III. EXPERIMENTS

A. Implementation Details

We use Inception V3 as the CN, while other CNN models such as VGG, ResNet can also be adopted. For all datasets, five landmarks are extracted, followed by face normalization. In training, input are randomly cropped from four corners or center. Random horizontal flip is also employed. Training list is shuffled at the beginning of each training epoch.

The CN is initialized using parameters pretrained on ImageNet, while others are randomly initialized. The training starts by tuning only FMG for 300 epochs, using Adam optimizer. The learning rate is initialized as 10^{-4} and decayed linearly to 0 from epoch 150. After that, the whole framework is jointly trained, with $\lambda_1 = 10$ and $\lambda_2 = 1$ in (5). The learning rate for FMG is reset to 10^{-5}, while the rest uses 10^{-4}. We jointly train the entire framework for another 200 epochs, and linearly decay the learning rates from epoch 100. The proposed model is implemented using PyTorch, and code is available at *https://github.com/donydchen/FMPN-FER*.

B. Expression Recognition Results

We consider two baseline methods: one is using only the classification network, referred as *CNN (baseline)*, the other is using the entire framework but without the training guidance from the ground truth masks, referred as *CNN (no l_G)*.

The **Extended Cohn-Kanade Dataset (CK+)** [9] is a laboratory-controlled benchmark dataset labelled with seven basic expressions. Following the settings of [8], the last three frames of each labelled sequence are extracted, results in a dataset with 981 images. These images are grouped according to the person identity and resort in an ascending order. And 10-fold subject-independent cross-validation experiments are conducted. As shown in TABLE I, our proposed method outperforms all other state-of-the-art approaches.Compared with the baselines, *CNN(baseline)* and *CNN(no l_G)*, our final model achieves a large gain, which indicates that the guidance from the moving muscle mask benefits FER. Fig. 3 (left) gives the details of the average accuracy in confusion matrix. We can see that *happiness* is the easiest one to be recognized, likely due to its unique feature of moving muscle around mouth (see Fig. 2), and *anger* has the relatively lowest recognition rate, which has some confusion with *disgust* and *sadness*.

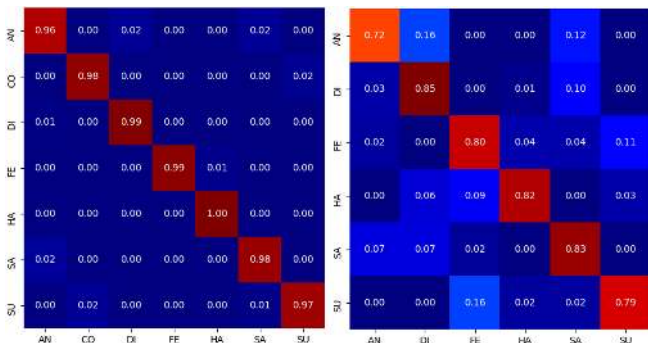| Approach	Setting	Accuracy (%)		
		CK+	MMI	AffectNet
LBP-TOP [11]	sequence-based	88.99	59.51	-
HOG 3D [12]	sequence-based	91.44	60.89	-
DTAGN(Joint) [3]	sequence-based	97.25	70.24	-
STM-Explet [13]	sequence-based	94.19	75.12	-
IACNN [14]	image-based	95.37	71.55	-
DeRF [8]	image-based	97.30	73.23	-
CNN (baseline)	image-based	90.78	68.81	60.86
CNN (no l_G)	image-based	91.82	63.10	60.01
FMPN (Ours)	image-based	**98.06**	**82.74**	**61.52**

Fig. 3. Confusion matrix on the CK+ (left) and MMI (right).

The **MMI** [10] is another laboratory-controlled dataset labelled with six basic expressions (without *contempt*). As a typical procedure, three peak frames around the center of each labelled sequence are selected, results in a dataset with a total of 624 expressive faces. Similar to CK+, 10-fold person-independent cross-validation experiments are conducted. As illustrated in TABLE I, compared with the image-based approaches, our method outperforms them by over 9.51%. Compared with the sequence-based methods, which use temporal information, our approach still achieves over 7.62% accuracy improvement. In addition, the large gaps between the two baseline models and our final model further demonstrate the importance of introducing the facial-motion mask and the usefulness of the guidance from the pre-computed ground-truth masks. Fig. 3 (right) gives the confusion matrix. We can see that it is similar to those of CK+. This indicates that facial expressions share similar patterns across different datasets.

The **AffectNet** [15] is a very large in-the-wild dataset. We conduct experiments on a subset of AffectNet. We randomly select around 3500 images for each of the seven basic expressions, resulting in a total of 24530 images. We randomly split it into training and testing set with a ratio of 9 to 1. Since the neutral faces corresponding to expressive faces are not available, we cannot generate the ground truth masks based on this dataset itself. Thus, we borrow the ground truth masks from CK+. Since AffectNet has enough data, we remove the pretraining for FMG. All other settings, including hyperparameters, learning rates, etc., remain unchanged.

As shown in TABLE I, transferring muscle moving masks from CK+ improves the recognition rate, which suggests that the information of facial moving muscles can be shared across different datasets, and it does help improve the performance of expression recognition. Note that it seems that the gain is not as significant as those in other datasets. This is mainly because of the large number of test images in AffectNet and the challenge of dealing with in-the-wild images. Other state-of-the-art methods did not report their results on AffectNet.

IV. CONCLUSION

We have proposed a novel FER framework, which incorporates prior knowledge. Particularly, for a given expressive face, we generate a facial mask to focus on facial muscle moving regions and we use the average differences between neutral faces and expressive faces as the guidance for the facial mask learning. Our method achieves the best results on CK+ and MMI datasets We have also reported our results on the large-scale in-the-wild dataset, AffectNet. These outstanding results demonstrates the effectiveness of the proposed model in FER.

REFERENCES

[1] B. Fasel and J. Luettin, "Automatic facial expression analysis: a survey," *PR*, vol. 36, no. 1, pp. 259–275, 2003.

[2] J. J. Bazzo and M. V. Lamar, "Recognizing facial actions using gabor wavelets with neutral face average difference," in *FG*, 2004, pp. 505–510.

[3] 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.

[4] N. Sun, Q. Li, R. Huan, J. Liu, and G. Han, "Deep spatial-temporal feature fusion for facial expression recognition in static images," *Pattern Recognition Letters*, 2017.

[5] M. Liu, S. Li, S. Shan, and X. Chen, "Au-inspired deep networks for facial expression feature learning," *Neurocomputing*, vol. 159, pp. 126–136, 2015.

[6] Z. Shao, Z. Liu, J. Cai, and L. Ma, "Deep adaptive attention for joint facial action unit detection and face alignment," in *ECCV*, 2018, pp. 705–720.

[7] Z. Shao, Z. Liu, J. Cai, Y. Wu, and L. Ma, "Facial action unit detection using attention and relation learning," *CoRR*, 2018.

[8] H. Yang, U. Ciftci, and L. Yin, "Facial expression recognition by de-expression residue learning," in *CVPR*, 2018, pp. 2168–2177.

[9] 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 *CVPRW*, 2010, pp. 94–101.

[10] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *ICME*, 2005, p. 5.

[11] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *TPAMI*, vol. 29, no. 6, pp. 915–928, 2007.

[12] A. Klaser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *BMVC*, 2008, pp. 275–1.

[13] 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.

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

[15] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *arXiv preprint arXiv:1708.03985*, 2017.