Toward Personalized Emotion Recognition: A Face Recognition Based Attention Method for Facial Emotion Recognition

Mostafa Shahabinejad¹, Yang Wang¹,², Yuanhao Yu¹, Jin Tang¹, Jiani Li¹
¹Noah’s Ark Lab, Huawei Technologies ²University of Manitoba, Canada

Abstract—This paper aims to address the subject-dependent challenge of the facial emotion recognition (FER) task. To accomplish this, we propose a novel face recognition based attention FER (FRA-FER) framework which propagates subtle face recognition (FR) features through the FER network. Particularly, first a spatial attention map from the feature maps of an FR convolutional neural network (CNN) is created and then it is fused into the FER-CNN. By doing this FR feature propagation, the FER network is personalized as it takes the advantage of the FR features learned from large-scale face recognition datasets. Experiments on the two challenging datasets AffectNet and AFEW demonstrate the superiority of our proposed FRA-FER network to the state-of-the-art work.

I. INTRODUCTION

Emotion recognition has a variety of applications, e.g., in human-computer interaction [1], [2], mental diseases diagnosis [3], avatars and computer animations [4], and more [5]. Several automatic emotion recognition methods based on speech [6], face [7], audio-visual [8], context [9], and multimodal sensor data [5] have been proposed. The focus of this work is on the expression recognition from facial features.

Convolutional neural network (CNN) based methods have made a lot of progress in automatic facial expression recognition (FER) [10], [11], [12], [13], [14], [15], [16], yet accurate facial expression recognition remains challenging because different individuals may have different ways of expressing the same emotion. There are many attempts in the FER literature to address the subjective notion of the emotion expression, e.g., see [17], [7], [18], [19], [20], [21] and references therein. To tackle the subject variation challenge, these methods use either a loss function to learn identity-related information [18], [22], [23] or generative adversarial networks (GANs) to generate expressions from facial images [20], [21]. The loss-based methods are dependent on the identity-related images and perform weakly when identities are limited, while the GAN-based methods depend largely on the unrealistic generated images for the classifier which result in a poorer expression recognition [17].

Unlike the previously proposed identity-robust FER methods, we consider fusing face recognition (FR) features into the FER network in order to personalize FER systems (Fig. 1). Because FR facial features learned from enormous identities from in-the-wild datasets contain subtle, rich information about individuals’ faces, we incorporate facial features from the FR-CNN and propagate them through the FER-CNN. More specifically, as shown in Fig. 1, a face image is passed through a trainable FER network and a frozen FR network. The features from the frozen FR-CNN are then used to create a spatial attention map which is later applied to the features of the FER-CNN. This leads to an improvement of the FER system with a negligible computation cost.

To the best of our knowledge, our work is the first to incorporate FR features to personalize the FER system. We show in the experiment section that this leads to a performance boost of 1.98% for the challenging AffectNet dataset [24] leading to the state-of-the-art results. Note that this performance improvement is achieved without increasing the capacity of the backbone network as the newly added FR-CNN parameters are frozen. It is also worth mentioning that the state-of-the-art results achieved by our method require only a minimal increase in computational cost compared to the backbone network because FR features are extracted from a very early layer of the FR-CNN.

The rest of the paper is organized as follows. In Section II, we summarize the related work. In Section III, we present details of our personalized FER networks. The experimental evaluation is conducted in Section IV. The conclusion is discussed in Section V of the paper.

II. RELATED WORK

Three general approaches can be considered in the research area of automatic facial emotion recognition. First, categorical models assign discrete emotion labels to the subject. Second, dimensional models describe emotions with two
(valence-arousal [25]) or three (pleasure-arousal-dominance [26]) continuous values. Third, facial action coding system (FACS) models detect the presence and the strength of different action units defined by FACS [27]. For recent surveys, please refer to [7], [28].

Here, we briefly review the literature that is most relevant to our work, namely categorical FER and FR. We first overview categorical FER. We then mention the FR state-of-the-art methods and the advantage of using them in the FER system.

A. Categorical FER

Categorical FER models assign emotion labels such as angry, happy, sad, etc. to the face image. The subject-dependent nature of the expression recognition makes the task challenging. There are several works addressing the subject-dependent issue of the FER which can be categorized into two general groups. The first group tries to learn identity along with expression through an identity loss via multi-task learning, while the second group tries to synthesize identity-based images through generative adversarial networks (GANs). In the following, we briefly review these methods.

In [18], an identity-sensitive contrastive loss is used to learn identity information. Their proposed method has two identical sub-networks, where one network is responsible to learn expression-discriminative features and the other one is responsible to learn identity-related features. The contrastive loss of this method suffers from drastic data expansion during task learning, while the second group tries to synthesize realistic generated images for the classifier which in turn results in a weak expression recognition.

Although all these approaches have shown great performance, they do not take into consideration the face recognition feature as a tool to satisfy the subjective-dependent nature of the facial expression problem. In Section IV, we show that propagating face recognition feature through the FER network leads to the state-of-the-art performance.

B. Face Recognition:

Face recognition (FR) is a problem related to FER. Current FR systems have achieved a very high accuracy. For example for the in-the-wild Megaface dataset which contains 1M images from 690K individuals with no constraint on expression, pose, and exposure [29], Arcface network [30] achieves an accuracy of %98.47. In this paper, we take advantage of the pretrained FR models learned from large-scale face recognition datasets to improve FER.

III. PROPOSED METHOD

In this section, we present our proposed personalized FR-based attention FER (FRA-FER) model. We first formally define the problem and overview the architecture of the FRA-FER in Section III.A. Then, we present details of the FRA-FER network in Section III.B.

A. Architecture Overview

Here, we overview the architecture of our proposed personalized FRA-FER CNN. Given a facial emotion recognition dataset $D = \{ (I_i, y_i) \}_{i=1}^{N}$ with $N$ images, where each image $I_i \in \mathcal{I} = \{ I_1, I_2, \cdots, I_N \}$ has a spatial resolution of $3 \times H \times W$ and is labeled from $C$ predefined emotion classes $y_i \in \mathcal{Y} = \{ 0, 1, \cdots, C-1 \}$, our goal is to personalize the FER classifier $\mathcal{F}_{FER}: \mathcal{I} \to \mathcal{Y}$ by propagating subtle FR features through the classifier. To this end, as depicted in Fig. 2, first the face image $I$ is passed through an $\mathcal{F}_{FR}$ network, where $\mathcal{F}_{FR}$ represents a frozen CNN which has learned face embedding through an FR database. Then a 3D feature map is extracted from some hidden layers of the FR-CNN. From the extracted 3D feature map, a 2D spatial attention map is constructed which is then applied to the feature map of a
trainable FER-CNN. Using this method, features of the FR-CNN are propagated through the FER-CNN which in turn improves the accuracy of the FER classifier. In the following, we present details of our methodology.

B. FRA-FER Network

Here, we present the details of our proposed personalized FRA-FER CNN. For a face image $I$, let

$$H_{FR_1, \ldots, FR_L} = \mathcal{F}_{FR}(I; \theta_{FR}),$$

where $\mathcal{F}_{FR}$ with the frozen parameters $\theta_{FR}$ is an FR-CNN with $L$ layers and $H_{FR_l}$ is the output of the $l$-th layer of the network with $l \in \{1, \ldots, L\}$. Similarly, let

$$H_{FER_1, \ldots, FER_{L'}} = \mathcal{F}_{FER}(I; \theta_{FER}),$$

where $\mathcal{F}_{FER}$ with trainable parameters $\theta_{FER}$ represents a FER classifier with $L'$ layers and $H_{FER_{l'}}$ is the output of the $l'$-th network layer with $l' \in \{1, \ldots, L'\}$.

As depicted in Fig. 3, in order to propagate the FR features through the FER networks, we create a spatial attention map from the $p$-th feature map of the FR-CNN and apply it to the $q$-th feature map of the FER-CNN as follows. Let $H_{FR_p} \in \mathbb{R}^{C_p \times H_p \times L_p}$ be the $p$-th feature map of FR-CNN, where $C_p$, $H_p$, and $W_p$ represent number of channels, height, and width of the feature map $H_{FR_p}$, respectively. Similarly, let $H_{FER_q} \in \mathbb{R}^{C_q \times H_q \times L_q}$ be the $q$-th feature map of FER-CNN, where $C_q$, $H_q$, and $W_q$ represent number of channels, height, and width of the feature map $H_{FER_q}$, respectively. First, the spatial attention map $S_{FER_q^p} \in \mathbb{R}^{1 \times H_q \times L_q}$ is generated as follows:

$$S_{FER_q^p} = \text{Interpolate} (\text{Sigmoid}(\text{Conv}(H_{FR_p}))),$$

where $\text{Conv}$, $\text{Sigmoid}$, and $\text{Interpolate}$ stand for a 2D convolution with $C_p$ kernels of size $1 \times 1$, the sigmoid function, and a 2D resizing operation, respectively. This spatial attention map is then applied channel-wise to the $q$-th feature map of the FER-CNN as follows:

$$\hat{H}_{FER_q^p} = \text{Mult}(S_{FER_q^p}, H_{FER_q}),$$

where $\text{Mult}$ represents the spatial channel-wise multiplication and $\hat{H}_{FER_q^p} \in \mathbb{R}^{C_q \times H_q \times L_q}$ represents the feature map $H_{FER_q}$ with the applied attention map. The final feature map $H_{FER_q^p} \in \mathbb{R}^{C_q \times H_q \times L_q}$ is obtained as

$$H_{FER_q^p} = \text{Max}(\hat{H}_{FER_q^p}, H_{FER_q}),$$

where $\text{Max}$ represents the element-wise maximum operation. Then the FER-CNN proceeds with $H_{FER_q^p}$ to generate FER features.

We note that the parameters $\theta_{FR}$ are learned offline using an FR databases and frozen in our FRA-FER network, while the parameters $\theta_{FER}$ are trainable. The pretrained FR models capture information about a person’s identity. By propagating the learned FR features through the FER network as shown in Eq. (4), we achieve a FER-CNN specific to each person.

Fig. 4 shows five image samples of the AffectNet dataset [24] along with their corresponding features. In this figure, the first and the second columns show the input image and the attention map, respectively. And the third and the forth columns show the spatial average of the FER feature map before and after applying the spatial attention map, respectively. As this figure shows, FER feature maps with the spatial attention (the forth column) provide more subtle features compared to the FER features with no spatial attention (the third column).

IV. EXPERIMENTS

In this section, we present our experiment results. To have a fair comparison, we adapt the EfficientNet-B0 network from the state-of-the-art work in [14] as the backbone network and follow their training and evaluation procedures.

For both FR and FER networks, we used the same EfficientNet-B0 network [31], [14]. We have found that getting the attention map from an early layer of the FR network and applying it to an early layer of the FER network
is more beneficial for the expression recognition task. Getting features from an early layer of the FR network has also the advantage of less computation. We get the $H_{FER^{att}}$ from (5) by setting $p$ to $1$ in (3) and apply it using (4) by setting $q$ to $1$. This results in a spatial attention map $S_{FR^{att}}$ of size $1 \times 112 \times 112$ and a FER feature map $H_{FER^{att}}$ of size $32 \times 112 \times 112$ for EfficientNet-B0 with the face image of size $3 \times 224 \times 224$. In our experiments, we use the same train parameters as in [14]: 6 epochs, batch of size 8, and $3 \times 10^{-5}$ flat learning rate.

We report accuracy of the validation sets corresponding to two challenging datasets AffectNet [24] and acted facial expression in-the-wild (AFEW) [32]. In the following, we first present our results on these two datasets and then show the ablation study on the AffectNet dataset.

### A. Results of AffectNet Dataset

AffectNet is the largest emotion recognition database containing more than 1M facial images gathered from the Internet. For the categorical emotion model of this dataset, there are more than 287K images for the train set and 8K images for the validation set (500 images per class) with 8 discrete classes: Anger, Contempt, Disgust, Fear, Happy, Neutral, Sad, and Surprise. Table I shows the validation accuracy results for the AffectNet dataset where our proposed personalized FRA-FER network achieves the state-of-the-art results for both 8 and 7 classes.

### B. Results of AFEW Dataset

The AFEW dataset contains 773 training and 383 validation samples collected from TV series and movies. There are 7 emotion classes corresponding with this dataset: Anger, Disgust, Fear, Happy, Neutral, Sad, Surprise. For the temporal feature extraction, we follow the same steps as in Section 3.3 of [14]. Table II shows the validation accuracy results for the AFEW dataset where our proposed personalized FRA-FER network achieves the state-of-the-art performance.

### C. Ablation Study

Here, we present our ablation study by reporting the accuracy of the validation set of the AffectNet dataset with 8 emotion classes. More specifically, we present results of no FR fusion, FR-FER feature concatenation and three self-attention methods of the spatial squeeze and channel excitation (SSCE) [35], channel squeeze and spatial excitation (CSSE) [34], and concurrent spatial and channel squeeze and spatial excitation (CSCSSE) [34]. As Table III shows, creating the attention map from the features of the FR network using our proposed method outperforms other fusion and self-attention methods.

### V. CONCLUSION

In this work, we have proposed a novel way of applying attention mechanism to the facial expression recognition convolutional neural network (FER-CNN). More specifically, rather than getting the attention map from the same FER-CNN, we propose to generate the attention map from a face recognition (FR) CNN. By doing this, we can make the FER network personalized by propagating the subtle FR features through the FER network. We have demonstrated that this approach results in the state-of-the-art performance for the two challenging datasets AffectNet and AFEW.
REFERENCES


[38] Albert Mehrabian. Basic dimensions for a general psychological theory: Implications for personality, social, environmental, and developmental studies. 1980.