Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/mbe

MBE, 20(6): 11101–11115.
DOI: 10.3934/mbe.2023491
Received: 19 January 2023
Revised: 27 March 2023
Accepted: 13 April 2023
Published: 24 April 2023

Research article

Leveraging ResNet and label distribution in advanced intelligent systems for facial expression recognition

Zhenggeng Qu[1,2,*] **and Danying Niu**[3]

[1] College of Mathematics and Computer Application, Shangluo University, Shaanxi 726000, China
[2] Engineering Research Center of Qinling Health Welfare Big Data, Shaanxi 726000, China
[3] Shangluo Central Hospital, Shaanxi 726000, China

* **Correspondence:** Email: quzhenggeng@163.com.

Abstract: With the development of AI (Artificial Intelligence), facial expression recognition (FER) is a hot topic in computer vision tasks. Many existing works employ a single label for FER. Therefore, the label distribution problem has not been considered for FER. In addition, some discriminative features can not be captured well. To overcome these problems, we propose a novel framework, ResFace, for FER. It has the following modules: 1) a local feature extraction module in which ResNet-18 and ResNet-50 are used to extract the local features for the following feature aggregation; 2) a channel feature aggregation module, in which a channel-spatial feature aggregation method is adopted to learn the high-level features for FER; 3) a compact feature aggregation module, in which several convolutional operations are used to learn the label distributions to interact with the softmax layer. Extensive experiments conducted on the FER+ and Real-world Affective Faces databases demonstrate that the proposed approach obtains comparable performances: 89.87% and 88.38%, respectively.

Keywords: affective computing; facial expression recognition; deep learning

1. Introduction

With the development of artificial intelligence, the field of facial expression recognition (FER) has attracted researchers from various fields, e.g., computer vision, affective computing, etc. Recently, deep learning (DL) technology has obtained excellent performance for FER [1, 2]. In general, the features can be classified as either hand-crafted or deep learned features. Examples of hand-crafting features include the use of the Gabor filter [3], the histogram of gradients [4], and local binary patterns [5]. But DL has the following advantages over hand-crafted features: 1) strong learning ability; 2) Many well-adaptive neural network layers with a wide width, which can theoretically map to any function to solve very complex problems; 3) Excellent portability [6]. [7] proposes a two-stage method

to fine-tune the face network for FER. Cai et al. [8] design a novel function named island loss to capture informative features for FER. [9] proposes a novel region attention architecture and adopts ensemble learning to learn discriminative features, which obtains state-of-the-art performance for FER.

Hence, the above method obtains a promising performance for FER. To lessen the computational cost for FER, some researchers try to design lightweight architectures to address the problem, e.g., MobileNet [10, 11], ShuffleNet [12, 13], and another deep models [14–20]. However, there are some challenges, such as occlusion and variations of the pose, which may lead to uninformative features being extracted, which makes for a poor performance for FER. Therefore, various researchers have tried to develop robust and lightweight architectures for FER. Nevertheless, there are still the following problems: 1) the above methods do not consider the label distribution; 2) the performance still needs to be improved in extracting discriminative features for FER.

Moreover, the psychologist's study [21] and existing FER work [22] have demonstrate that the most emotions occur as combinations, mixtures, or compounds of the basic emotions, and multiple emotions always have different intensities from a single facial image, especially in the real world. Therefore, to further promote the performance of the FER model, training the FER models by label distribution called label distribution learning (LDL) instead of a single label seems more reasonable. Moreover, there are various researches have shown that the LDL can also overcome the noise problem caused by the subjectiveness of annotators and ambiguous in facial images [23]. Different from these approaches, we introduce a simple but robust model by training a ResNet and label distribution generator (LDG) to generate the label distribution directly.

In this paper, to solve the aforementioned problems, we introduce a novel architecture to model the discriminative features for FER. Specifically, the proposed architecture includes the following main components: 1) local feature extraction. Based on experience, ResNet-50 [24] is adopted to extract the local low-level features. Meanwhile, to further know the efficiency of the proposed architecture, other ResNet variants have also been evaluated. 2) To model the discriminative features based on the local features, a feature aggregation module has been designed. Specifically, convolutional and global average pooling (GAP) operation are adopted to transform and aggregate the local features for the next stage. 3) To treat the label distribution problem, a cascade of convolutional operations are also adopted to model the label distribution for FER. Meanwhile, a label distribution module and the main architecture are also interactive to obtain the final performance of FER. In this paper, we focus on the local and global feature extraction stages for FER.

1.1. Contributions

The main contributions can be summarized as follows:

1) A novel architecture, ResFace, is designed. In the architecture, we focus on extracting the local and global discriminative features from the static images for FER. We then aggregate the local features so as to reduce the useless features.

2) We explore and consider the label distribution in this task. This can overcome the noise issue for FER. In addition, it also can improve the performance. In comparison with the modelling using only a single label, considering the label distribution can promote the performance for FER.

3) Extensive experiments have been conducted on the RAF-DB and FER2013 databases, and

promising performance was obtained. In addition, the effectiveness of the proposed method has been evaluated on these databases.

The remainder of this paper is structured as follows. We briefly discuss the existing work based on static images for FER in Section 2. Then we introduce the proposed method in Section 3. Subsection 4.1 introduces the used databases and presents the experimental results. Conclusions and future work are described in Section 5.

2. Related works

Based on DL, many researchers have focused on designing various shallow and deep architectures for FER. In particular, the emergence of transformer has brought about great success in many fields. Therefore, we briefly divide the previous work into three aspects: hand-crafted features, CNN-based features, and transformer-based features for FER.

Firstly, we briefly discuss the hand-crafted features and CNN-based features for FER. In [25], the authors propose a hybrid architecture for FER. Specifically, the architecture includes dense and regular Scale Invariant Feature Transform (SIFT) features for FER. They evaluate the proposed method on the FER2013 and CK+ databases, and obtain a performance of 73.4% on FER2013 and 99.1% on CK+.

In [26], the authors design a two-stream framework with a spatial and temporal convolutional neural network (CNN) and local binary patterns for three orthogonal planes (LBP-TOP) feature for FER. Specifically, the framework considers the spatial and temporal patterns in the expression and non-expression frame. The authors validate the proposed method method on CK+, and obtain a comparable performance.

Shao et al. [27] proposes three architectures: a shallow network Light-CNN, a dual-branch CNN, and a pre-trained CNN. For the shallow light-CNN, a fully convolutional neural network includes six depthwise separable residual convolution modules to overcome the problem of complex topology and over-fitting. The dual-branch CNN can extract deep learning features and LBP features at the same time. The pre-trained CNN uses transfer learning technology to cope with the small amount of training data. Extensive experiments were conducted on the CK+, BU-3DEF and FER2013 datasets, and demonstrate the effectiveness of the proposed method when compared to the state of the art methods.

[23] introduces a novel architecture label distribution learning on auxiliary label space graphs (LDL-ALSG) for FER. Specifically, the framework uses topological information to model the facial features from action units and facial landmarks. They assume that facial images should have similar expression distributions to their neighbours in the label space of action unit recognition and facial landmark detection. The proposed method is validated on various datasets and obtained promising performance when compared the state of the art methods.

In [28], the authors propose an informative and discriminative Feature Learning (IDFL) framework. Two components make up the framework: a multi-Path Attention Convolutional Neural Network (MPACNN) and Balanced Separate loss (BS loss), for both basic and compound high-accuracy FER in the wild. Specifically, MPACNN adopts many paths to capture discriminative features. In addition, the BS loss is adopted as the objective function of MPACNN, so the model can capture informative and discriminative features simultaneously, obtaining competitive results. The authors evaluated the proposed method on seven databases, and the performances demonstrate that

that approach obtains state-of-the-art results on both basic and compound expressions.

[29] proposes a novel architecture for FER, in which deep sparse autoencoders (DSAE) is adopted to extract informative features from the data. Extensive experiments were conducted on the CK+ database, obtaining an accuracy of 95.79%. In addition, that method was also used to recognize eight facial expressions, obtaining a comparable accuracy when compared with most other methods.

[30] designs a dynamic multi-channel metric learning network (DML-Net) to reduce the effects of pose and identity for FER. Specifically, DML-Net can learn the local and global features from the facial regions. DML-Net is an end-to-end trainable architecture to model discriminative features for FER. Extensive experiments obtained accuracies of 88.2% on KDEF, 83.5% on BU-3DFE, 93.5% on Multi-PIE, and 54.36% on SFEW.

In [31], Zhang et al. introduced a novel scheme to model the problem for FER. The scheme, referred to as relative uncertainty learning (RUL), considers these uncertainties as weights to mix facial features and design an add-up loss to encourage uncertainty learning. Extensive experiments show that RUL outperforms state-of-the-art FER methods.

[32] proposes a deep attentive center loss (DACL) method to learn the discriminative features to improve the informative features. Specifically, the DACL combines an attention mechanism with spatial feature maps to learn the discriminative features for FER. The estimated weights accommodate the sparse formulation of center loss to selectively achieve intra-class compactness and inter-class separation for the relevant information in the embedding space. They evaluate the proposed method on the public databases and demonstrate its effectiveness.

[33] proposes a disentangled facial expression recognition (IPD-FER) model for FER. The authors consider the ensemble facial expressions, such as identity, pose, and expression. Therefore, three different encoders are designed: an identity encoder, a pose encoder, and an expression encoder. Experiments validate its capacity on both lab-controlled and in-the-wild databases and they obtain state-of-the-art recognition performance.

[34] proposes a coarse-to-fine cascaded network with smooth predicting (CFC-SP) for FER. The CFC-SP includes two models: coarse-to-fine cascaded networks (CFC) and smooth predicting (SP). The proposed architecture can model the discriminative features for FER. They got 3rd place in the expression classification challenge of the 3rd competition on affective behavior analysis in-the-wild.

In [35], the authors propose a new feature decomposition and reconstruction learning (FDRL) method for FER. The framework includes two architectures: a feature decomposition network (FDN) and a feature reconstruction network (FRN). They obtain promising performance both on the lab controlled databases (including CK+, MMI, and Oulu-CASIA) and in-the-wild databases (including RAF-DB and SFEW).

Based on the work reviewed above, we can conclude that: 1) the existing work only focuses on learning the discriminative features in static images, and use different network structures to mine fine-grained emotional feature embedding; 2) the existing work mainly focused on adopting a single label for FER. Therefore, we have designed a novel framework to address these problems.

2.1. Architecture overview

To learn the discriminative features using the hybrid architecture, a brief overview will now be presented. Figure 1 presents the architecture. The pipeline consists of the following modules: 1) a local feature extraction module in which ResNet (ResNet-18, ResNet-50) are used to extract the local

features for the following feature aggregation; 2) a channel feature aggregation module, in which a channel-spatial feature aggregation method is adopted to learn the high-level features; 3) a compact feature aggregation module, in which several convolutional operation are used to learn the label distributions to interact with the softmax layer. In the following, we provide a detailed introduction to each module of the architecture. The remainder of this section is structured as follows. Subsection 3.1 details the local feature extraction module, the channel-spatial feature aggregation module is presented in Subsection 3.2, and then Subsection 3.3 discusses the label distribution generator. The pipeline of the proposed approach is illustrated in Figure 1.

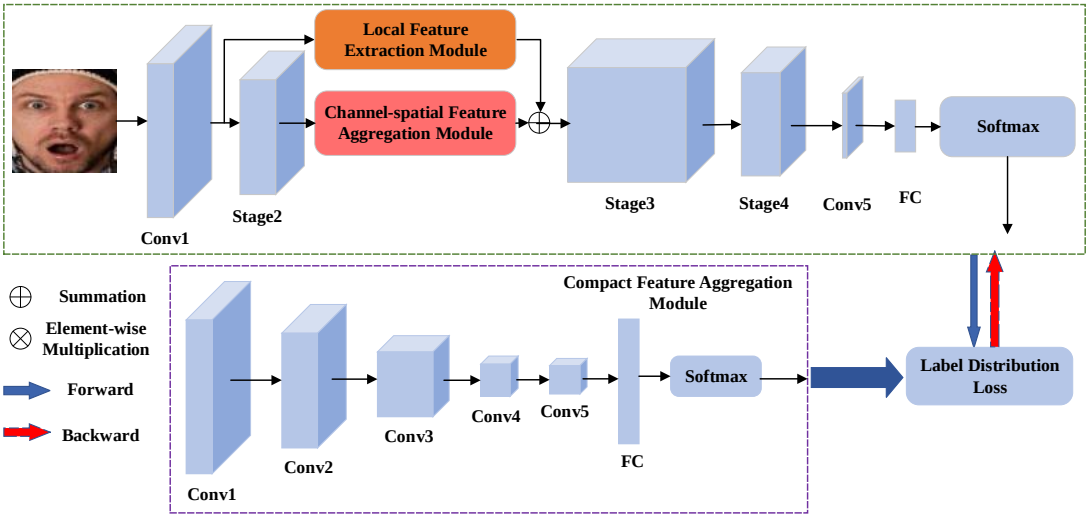

Figure 1. Illustration of the proposed pipeline for ADE. The pipeline consists of the following modules: 1) local feature extraction module, 2) channel feature aggregation module, 3) compact feature aggregation module. The local feature extraction module learns the local deep features using ResNet-18 and ResNet-50. The channel feature aggregation module can be used to learn the high-level features. In the compact feature aggregation module, several convolutional operation are adopted for feature learning to interact with the label distribution learning for the final FER.

3. Our architecture

3.1. Local feature extraction

By now, deep learning technology has been adopted as an efficient feature extraction in computer vision. DCNN captures local features with a hierarchical pattern by convolutional operations and maintains the local patterns as feature maps. Meanwhile, ViT is also proposed to aggregate global patterns from the compressed patch embedding by a soft pattern with the cascaded self-attention modules. In this work, the ResNet (ResNet-18, ResNet-50) architecture is used to learn the discriminative deep features.

Due to the small amount of training data, the pre-trained model can be divided in two ways for FER. First, we explore the existing pre-trained deep models and adopt the discriminative models for FER.

The other way is that the ensemble architecture and its sub-architecture are fine-tuned to fit our task. Therefore, the weights of the deep model are taken as the values of the parameters for the new task and updated in the training stage.

As for the FER task, to extract the discriminative features, ResNet (ResNet-18, ResNet-50) are adopted to extract the deep features, followed by adopting different modules to extract the mid-level and high-level features for FER. In our task, to make a clear explanation for the FER (Figure 3), we only use ResNet-18 as an example. As illustrated in Figure 2, ResNet-18 contains 16 convolution layers, 2 downsampling layers, and 2 fully connected layers. In ResNet-18, the convolutional layer is 224×224 with a kernel size of 7×7. As is known, the kernel size of the other layers is 3×3. Followed by the last convolutional layer, an eigenvector is obtained by full connection. ResNet-50 has a similar architecture to that of ResNet-18: the only difference is that it contains more layers.

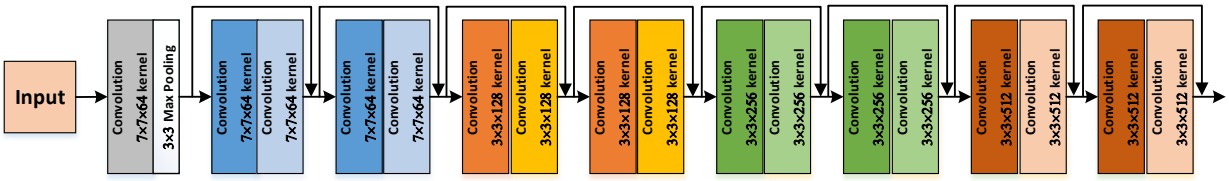

Figure 2. Illustration of the ResNet-18 architecture. ResNet-18 include 17 convolutional layers and 1 fully connected layer.

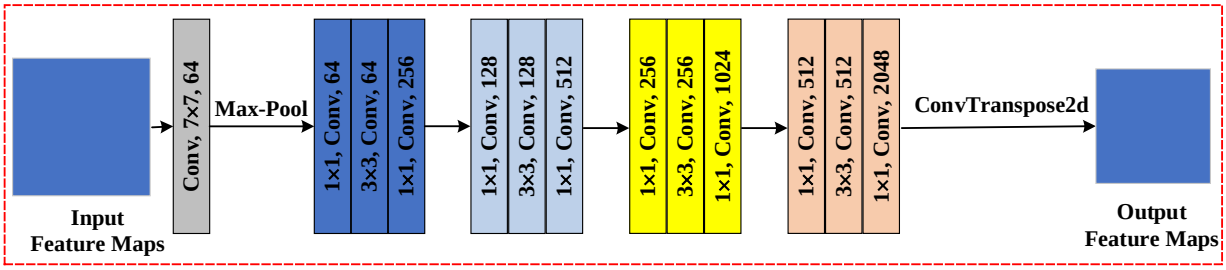

Figure 3. Illustration of the local feature extraction. After learning the compact features by the variants of ResNet, ConvTranspose2d operation is used to augment and upsample the size of the feature maps to make the features for the following feature learning and aggregation.

After the feature extraction by the ResNet variants, discriminative feature representations are obtained from the architectures. More importantly, to improve the efficiency of the FER model, we use the novel architecture ShuffleNet-V2 [13], which includes Conv1, Stages 2–4, and Conv5, and is adopted as the main backbone network in our task. Meanwhile, to overcome the issues of occlusion and pose variation in real-world scenes, two modules are adopted: a local-feature extraction module and a channel-spatial feature aggregation module. In addition, a novel label distribution learning method is introduced [21].

Existing work has suggested that local facial features are beneficial to FER [9, 36]. However, the above mentioned methods mainly focused on the facial landmarks, and do not consider the local facial features. Therefore, we adopted ResNet to extract local region features, to provide an solid foundation

for the following global feature representation learning. The architecture is shown in Figure 3 (dashed red rectangle). Assume that the input image has a size of $224 \times 224 \times 3$. After performing the first convolutional (conv1) operation, the feature map has a size of $56 \times 56 \times 29$. To make a convenient representation, let us assume the output is $Fea_{conv1} \in \mathbb{R}^{H \times W \times C}$. After inputting the feature map Fea_{conv1} into the ResNet architecture, a new feature map $Fea_{local} \in \mathbb{R}^{H' \times W' \times C'}$ is obtained.

3.2. Compact feature aggregation

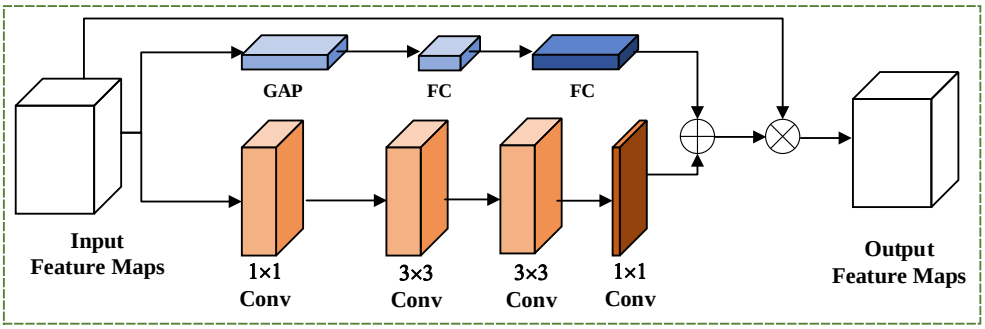

Figure 4. Illustration of the channel-spatial aggregation module for feature learning. To learn the compact features, the first branch consists of one GAP layer and two FC layers, and the second branch consists of two 1×1 layers and two 3×3 layers. After the following operation, the learned features are concatenated and multiplied with the original feature maps to generate the output feature maps. $\oplus$ denotes summation. $\otimes$ denotes element-wise multiplication.

After the local feature extraction, discriminative features are obtained. Although ResNet has the capacity to learn informative features, there are also some redundant features in it. Therefore, we adopt a channel-spatial aggregation method to aggregate the representative features after Stage 2. In our work, we change the channel-spatial aggregation method based on [37]. The framework of the channel-spatial feature aggregation module is illustrated in Figure 4 (green dashed rectangle). After the feature maps are input into Stage 2, the high-level feature maps $(Fea_{stage2}) \in \mathbb{R}^{H' \times W' \times C'}$ are generated. In this stage, the procedure can be split into two sub-procedures for feature aggregation. The channel heatmaps $Z_{channel}(Fea_{stage2}) \in \mathbb{R}^{C'}$ and the spatial heatmaps $Z_{spatial}(Fea_{stage2}) \in {}^{H' \times W'}$ are calculated by two parallele procedures. After that, all the heatmaps $Z(Fea_{stage2})$ can be written as

$$Z(Fea_{stage2}) = \sigma(Z_{channel}(Fea_{stage2}) + Z_{spatial}(Fea_{stage2})) \tag{3.1}$$

where σ denotes a sigmoid function. The channel heatmaps $Z_{channel}(Fea_{stage2}) \in \mathbb{R}^{C'}$ and the spatial heatmaps $Z_{spatial}(Fea_{stage2}) \in \mathbb{R}^{H' \times W'}$ are resized to $\mathbb{R}^{H' \times W' \times C'}$ before performing the addition operation. Lastly, the final global feature maps are formed as follows:

$$Fea_{modulated} = (Fea_{stage2}) \otimes Z(Fea_{stage2}). \tag{3.2}$$

When we perform the local feature extraction and bottleneck attention feature aggregation, the obtained local–global features can be written as

$$Fea_{final} = Fea_{local} + Fea_{modulated}. \tag{3.3}$$

After performing the feature extraction and aggregation, the architecture is able to learn local features and global-salience. In our work, to learn the discriminative features, we only adopt the local feature extraction and channel-spatial feature aggregation after Stage 2. The reason is that the feature maps contain the informative features by combining the local and channel-spatial features.

3.3. Label distribution learning

For the FER task, the annotation of the facial images is often difficult to recognize. In view of this, a label distribution generator (LDG) has been designed to generate the label distribution for the deep model training. The architecture is illustrated in Figure 1 (the pink dashed rectangle). To ensure the capacity of the deep models, the LDG is fixed in the training stage.

Suppose given a facial image X, and the corresponding label $la \in \{0, 1, ..., l-1\}$, where l is the number of categories of expressions. The function of the LDG is to generate a distribution $Dis = (Dis_0, Dis_1, ..., Dis_{l-1})$, where $\sum_{j=0}^{l-1} \phi_j = 1$. A feature vector $Fv = (Fv_0, Fv_1, Fv_2, ..., Fv_{l-1})$ is generated. Then Dis can be calculated by a softmax function:

$$Dis = \frac{\exp(Fv_m)}{\sum_{n=0}^{l-1} \exp(Fv_n)} \tag{3.4}$$

where $m \in \{0, 1, ..., l-1\}$.

As is known, the existing FER methods also use a single label as ground truth. In our work, we use a label distribution as ground truth for FER. Cross-entropy is adopted to compute the distance between the predicted value and the output of LDG. Therefore, the label distribution loss can be written as

$$Loss_{all} = -\frac{1}{N \times l} \sum_{i=0}^{N-1} \sum_{m=0}^{l-1} dis_m^i \log(\overline{dis}_m^i) \tag{3.5}$$

where N is the number of samples, $\overline{dis} = (\overline{dis}_1, \overline{dis}_2, ..., \overline{dis}_{l-1}$ is the predicted label distribution, and i and j indicate the sample and the expression, respectively.

4. Experiments

In this part, we present the experimental validation of the introduced framework for FER. We first present the used datasets, and the experimental settings and evaluation standard. Finally, we discuss the experimental results for FER.

4.1. Datasets

RAF-DB [38] contains 30,000 facial images annotated with basic or compound expressions by 40 trained human coders. In order to make a fair comparison with the existing work, seven basic emotions, i.e., neutral, happiness, sadness, surprise, fear, disgust, and anger, are adopted in our experiment. The datasets consist of 12,271 images for training and 3068 images for testing.

FERPlus is an extension of FER2013 [39] in the the ICML 2013 Challenges. It is a large-scale and real-world dataset collected by the Google search engine, and consists of 28,709 training images, 3589 validation images, and 3589 test images. All the facial images are resized to 48×48.

4.2. Implementation details

In our experiment, we fine-tune the ResNet variants for FER. All the facial images are resized to 224×224. To overcome the over-fitting problem, random cropping and random horizontal flipping are adopted. The proposed architecture is trained from scratch. LDG is pre-trained on the face recognition dataset MS-Celeb-1M [40]. For LDG, the 50-layer Residual Network is used as the backbone network. For ResFace, the parameters were optimized via the SGD optimizer with an initial learning rate of 0.1 and a mini-batch size of 64. The Pytorch platform is used to train the deep models with the NVIDIA Titan RTX.

4.3. Results

In this part, we first discuss the results for FER. Then we compare them with state of the art methods for FER.

To validate the efficiency of the proposed method, we combine several architectures to learn the discriminative features for FER. Resnet-18 and ResNet-50 are used to learn the local features, and channel-spatial feature aggregation is used to learn the global features for the LDG for FER.

Table 1. Performance of ResNet-18 on the RAF-DB dataset.

Dataset	Architecture	# Params (M)	# MFLOPs	Accuracy (%)
RAF-DB	A	10.1	1642.86	88.00
	A + B	10.2	1659.13	88.02
	A + C	10.3	1675.39	88.07
	A + B + C	10.4	1691.66	88.21

Table 2. Performance of ResNet-50 on the RAF-DB dataset.

Dataset	Architecture	# Params (M)	# MFLOPs	Accuracy (%)
RAF-DB	A	12.1	1968.18	88.02
	A + B	12.2	1984.45	88.11
	A + C	12.3	2000.72	88.16
	A + B + C	12.4	2016.74	88.38

Table 3. Performance of ResNet-18 on the FER+ dataset.

Dataset	Architecture	# Params (M)	# MFLOPs	Accuracy (%)
FER+	A	10.1	1642.86	87.78
	A + B	10.2	1659.13	87.80
	A + C	10.3	1675.39	87.81
	A + B + C	10.4	1691.66	87.89

Table 4. Performance of ResNet-50 on the FER+ dataset.

Dataset	Architecture	# Params (M)	# MFLOPs	Accuracy (%)
FER+	A	12.1	1968.18	88.83
	A + B	12.2	1984.45	89.46
	A + C	12.3	2000.72	89.54
	A + B + C	12.4	2016.74	89.87

To make a simple explanation for the experimental results, we abbreviate the "Baseline", "Local-Feature Extractor" and "Channel-Spatial Modulator" as A, se B and C, respectively. To validate the effectiveness of the proposed method, we conductd extensive experiments with ResNet-18 and ResNet-50 on the RAF-DB and FER+ datasets for FER. From Table 1, one can note that "A + B + C" obtains the best performance, 88.21%. However, more parameters will be obtained, i.e., the parameters (M) of 10.4 and the MFLOPs with 1691.66. Meanwhile, for the ResNet-50 architecture, we obtain a comparable performance for the FER with an accuracy of 88.38% (see Table 2).

On the FER+ dataset, we also validate the capacity with the ResNet-18 and ResNet-50 architecture. From Table 3, one can note that "A + B + C" still obtain the best performance, with an accuracy of 54.89% for FER with the ResNet-18 architecture. To further validate the effectiveness of the proposed method, we also use ResNet-50 as a backbone to learn the discriminative features for FER (see Table 4).

4.3.1. Comparison with the state-of-the-art methods

To further compare the proposed method with the existing methods, we list several papers based on the DL technology for FER.

Table 5. Performance of the proposed method on the RAF-DB dataset.

Methods	# Params (M)	# MFLOPs	Accuracy (%)
IPA2LT [29]	> 23.52	> 4109.48	86.77
Separate-Loss [41]	11.18	1818.56	86.38
gACNN [36]	> 134.29	> 15479.79	85.07
RAN [9]	11.19	14548.45	86.90
LDL-ALSG [23]	23.52	4109.48	85.53
SCN [42]	11.18	1818.56	87.03
EfficientFace [43]	1.28	154.18	88.36
Ours ($ResNet-18_{(cross-entropy-loss)}$)	12.4	2156.12	83.24
Ours ($ResNet-50_{(cross-entropy-loss)}$)	13.1	2789.74	84.12
Ours (ResNet-18)	10.4	1691.66	88.21
Ours (ResNet-50)	12.4	2016.74	88.38

Table 6. Performance of the proposed method on the FER+ dataset.

Methods	Year	Accuracy (%)
VGG13 [44]	2016	85.10
ResNet-18+VGG16 [45]	2017	87.40
SENet-50 [46]	2018	88.80
RAN+ResNet-18 [9]	2020	88.55
RAN+VGG16 [9]	2020	89.16
SCN+ResNet-18 [42]	2020	88.01
SCN+ResNet-18+ArcFace [42]	2020	89.35
Ours (ResNet-18_(cross-entropy-loss))	2022	84.07
Ours (ResNet-50_(cross-entropy-loss))	2022	85.20
Ours (ResNet-18)	2022	87.89
Ours (ResNet-50)	2022	89.87

In this part, our goal is to compare some state-of-the-art FER methods applied to the RAF-DB and FER2013 databases. To obtain the best performance for FER, single label distribution is used by many methods. In Table 5, we show the results of state of art methods on RAF-DB for FER. Our ResFace (ResNet-18) achieves a comparable accuracy of 88.21%, and the ResNet-50 obtains a comparable accuracy of 88.38%. The results show that the proposed method can better adopt both local and global features, and the performances are advanced. Moreover, the comparison with other methods also shows the effectiveness of the proposed method.

For FER+, as shown in Table 6, the proposed method obtains a comparable accuracy of 87.89% and 89.87% with the ResNet-18 and ResNet-50, respectively. However, our method adopts the local, global and label distribution information for FER.

On the considered RAF-DB and FER+ datasets, our method has achieved comparable performance for FER. The advantage of our proposed method is that it can learn the informanative features for FER. Specifically, the first reason is that the variants of ResNet can learn the local features to reduce the noise of the features. Second, LDG can learn the distribution of labels, which can learn the informinative information for FER.

5. Conclusions

Nowadays, various deep learning methods are proposed for FER. However, they mainly focused on single label for this task. In this paper we proposed a novel architecture, referred to as ResFace, for facial expression recognition (FER). In this framework, we use three modules: a local feature extraction module, a channel-spatial feature aggregation module, and a compact feature aggregation module, as well as several convolutional operations to learn the label distributions to interact with the softmax layer. Extensive experiments conducted on the RAF-DB and FER+ datasets showed that it obtained promising performance for FER.

In the future, we will learn the local deep features, and capture the discriminative features for FER. In addition, we will leverage the label distribution learning scheme for FER.

Acknowledgments

This work was supported by the National Social Science Foundation of China (Grant No. 21XJY015), the Shaanxi Provincial Natural Science Foundation (Grant No. 2022JM-339), and the Shangluo Municipal Science and Technology Bureau (Grant No. 2020-Z-0045), the 2023 Annual Special Project on Philosophical and Social Science Research in Shaanxi Province (Grant No. 2023QN0233).

Conflict of interest

The authors declare there is no conflict of interest.

References

1. G. M. Jacob, B. Stenger, Facial action unit detection with transformers, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2021), 7680–7689. http://doi.org/10.1109/CVPR46437.2021.00759
2. X. Liu, L. Jin, X. Han, J. Lu, J. You, L. Kong, Identity-aware facial expression recognition in compressed video, *arXiv preprint*, (2021), arXiv:2101.00317. https://doi.org/10.48550/arXiv.2101.00317
3. C. Liu, H. Wechsler, Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition, *IEEE Trans. Image Process.*, **11** (2002), 467–476. https://doi.org/10.1109/TIP.2002.999679
4. N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, (2005), 886–893. https://doi.org/10.1109/CVPR.2005.177
5. C. Shan, S. Gong, P. W. McOwan, Facial expression recognition based on local binary patterns: A comprehensive study, *Image Vision Comput.*, **27** (2009), 803–816. https://doi.org/10.1016/j.imavis.2008.08.005
6. Y. LeCun, Y. Bengio, G. Hinton, Deep learning, *Nature*, **521** (2015), 436–444. https://doi.org/10.1038/nature14539
7. H. Ding, S. K. Zhou, R. Chellappa, Facenet2expnet: Regularizing a deep face recognition net for expression recognition, in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, (2017), 118–126. https://doi.org/10.1109/FG.2017.23
8. J. Cai, Z. Meng, A. S. Khan, Z. Li, J. O'Reilly, Y. Tong, Island loss for learning discriminative features in facial expression recognition, in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, (2018), 302–309. http://doi.org/10.1109/FG.2018.00051
9. K. Wang, X. Peng, J. Yang, D. Meng, Y. Qiao, Region attention networks for pose and occlusion robust facial expression recognition, *IEEE Trans. Image Process.*, **29** (2020), 4057–4069. http://doi.org/10.1109/TIP.2019.2956143

10. A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, et al., Mobilenets: Efficient convolutional neural networks for mobile vision applications, *arXiv preprint*, (2017), arXiv:1704.04861. https://doi.org/10.48550/arXiv.1704.04861
11. M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L. C. Chen, Mobilenetv2: Inverted residuals and linear bottlenecks, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2018), 4510–4520. https://doi.org/10.1109/CVPR.2018.00474
12. X. Zhang, X. Zhou, M. Lin, J. Sun, Shufflenet: An extremely efficient convolutional neural network for mobile devices, *arXiv preprint*, (2018), arXiv:1707.01083. https://doi.org/10.48550/arXiv.1707.01083
13. N. Ma, X. Zhang, H. T. Zheng, J. Sun, Shufflenet v2: Practical guidelines for efficient cnn architecture design, *arXiv preprint*, (2018), arXiv:1807.11164. https://doi.org/10.48550/arXiv.1807.11164
14. L. Yang, Y. Li, S. X. Yang, Y. Lu, T. Guo, K. Yu, Generative adversarial learning for intelligent trust management in 6G wireless networks, *arXiv preprint*, (2022), arXiv:2208.01221. https://doi.org/10.48550/arXiv.2208.01221
15. S. Xia, Z. Yao, G. Wu, Y. Li, Distributed offloading for cooperative intelligent transportation under heterogeneous networks, *IEEE Trans. Intell. Transp. Syst.*, **23** (2022), 16701–16714. http://doi.org/10.1109/TITS.2022.3190280
16. D. Peng, D. He, Y. Li, Z. Wang, Integrating terrestrial and satellite multibeam systems toward 6G: Techniques and challenges for interference mitigation, *IEEE Wireless Commun.*, **29** (2022), 24–31. http://doi.org/10.1109/MWC.002.00293
17. H. Li, M. Zhang, D. Chen, J. Zhang, M. Yang, Z. Li, Image color rendering based on hinge-cross-entropy gan in internet of medical things, *Comput. Model. Eng. Sci.*, **135** (2023), 779–794. https://doi.org/10.32604/cmes.2022.022369
18. J. Zhang, Q. Yan, X. Zhu, K. Yu, Smart industrial IoT empowered crowd sensing for safety monitoring in coal mine, *Digital Commun. Networks*, 2022, in press. https://doi.org/10.1016/j.dcan.2022.08.002
19. L. Huang, R. Nan, K. Chi, Q. Hua, K. Yu, N. Kumar, et al., Throughput guarantees for multi-cell wireless powered communication networks with non-orthogonal multiple access, *IEEE Trans. Veh. Technol.*, **71** (2022), 12104–12116. https://doi.org/10.1109/TVT.2022.3189699
20. H. Li, L. Hu, J. Zhang, Irregular mask image inpainting based on progressive generative adversarial networks, *Imaging Sci. J.*, **2023** (2023), 1–14. https://doi.org/10.1080/13682199.2023.2180834
21. R. Plutchik, A general psychoevolutionary theory of emotion, *Social Sci. Inf.*, **21** (1982), 529–553. https://doi.org/10.1177/053901882021004003
22. S. H. Gao, M. M. Cheng, K. Zhao, X. Y. Zhang, M. H. Yang, P. Torr, Res2net: A new multi-scale backbone architecture, *arXiv preprint*, (2019), arXiv:1904.01169. https://doi.org/10.48550/arXiv.1904.01169

23. S. Chen, J. Wang, Y. Chen, Z. Shi, X. Geng, Y. Rui, Label distribution learning on auxiliary label space graphs for facial expression recognition, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2020), 13984–13993. http://doi.org/10.1109/CVPR42600.2020.01400

24. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), 770–778. https://doi.org/10.1109/CVPR.2016.90

25. T. Connie, M. Al-Shabi, W. P. Cheah, M. Goh, Facial expression recognition using a hybrid cnn–sift aggregator, in *Multi-disciplinary Trends in Artificial Intelligence*, Springer, (2017), 139–149.

26. D. Feng, F. Ren, Dynamic facial expression recognition based on two-stream-cnn with LBP-TOP, in *2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS)*, (2018), 355–359. https://doi.org/10.1109/CCIS.2018.8691380

27. J. Shao, Y. Qian, Three convolutional neural network models for facial expression recognition in the wild, *Neurocomputing*, **355** (2019), 82–92. https://doi.org/10.1016/j.neucom.2019.05.005

28. Y. Li, Y. Lu, B. Chen, Z. Zhang, J. Li, G. Lu, et al., Learning informative and discriminative features for facial expression recognition in the wild, *IEEE Trans. Circuits Syst. Video Technol.*, **32** (2022), 3178–3189. https://doi.org/10.1109/TCSVT.2021.3103760

29. N. Zeng, H. Zhang, B. Song, W. Liu, Y. Li, A. M. Dobaie, Facial expression recognition via learning deep sparse autoencoders, *Neurocomputing*, **273** (2018), 643–649. https://doi.org/10.1016/j.neucom.2017.08.043

30. Y. Liu, W. Dai, F. Fang, Y. Chen, R. Huang, R. Wang, et al., Dynamic multi-channel metric network for joint pose-aware and identity-invariant facial expression recognition, *Inf. Sci.*, **578** (2021), 195–213. https://doi.org/10.1016/j.ins.2021.07.034

31. Y. Zhang, C. Wang, W. Deng, Relative uncertainty learning for facial expression recognition, in *35th Conference on Neural Information Processing Systems (NeurIPS 2021)*, (2021), 17616–17627.

32. A. H. Farzaneh, X. Qi, Facial expression recognition in the wild via deep attentive center loss, in 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), (2021), 2402–2411. https://doi.org/10.1109/WACV48630.2021.00245

33. J. Jiang, W. Deng, Disentangling identity and pose for facial expression recognition, *IEEE Trans. Affective Comput.*, **13** (2022), 1868–1878. https://doi.org/10.1109/TAFFC.2022.3197761

34. F. Xue, Z. Tan, Y. Zhu, Z. Ma, G. Guo, Coarse-to-fine cascaded networks with smooth predicting for video facial expression recognition, in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, (2022), 2412–2418. https://doi.org/10.1109/CVPRW56347.2022.00269

35. D. Ruan, Y. Yan, S. Lai, Z. Chai, C. Shen, H. Wang, Feature decomposition and reconstruction learning for effective facial expression recognition, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2021), 7660–7669.

36. Y. Li, J. Zeng, S. Shan, X. Chen, Occlusion aware facial expression recognition using cnn with attention mechanism, *IEEE Trans. Image Process.*, **28** (2019), 2439–2450. https://doi.org/10.1109/TIP.2018.2886767

37. J. Park, S. Woo, J. Y. Lee, I. S. Kweon, Bam: Bottleneck attention module, *arXiv preprint*, (2018), arXiv:1807.06514. https://doi.org/10.48550/arXiv.1807.06514

38. S. Li, W. Deng, J. Du, Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2017), 2852–2861.

39. I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, et al., Challenges in representation learning: A report on three machine learning contests, *Neural Networks*, **64** (2015), 59–63. https://doi.org/10.1016/j.neunet.2014.09.005

40. Y. Guo, L. Zhang, Y. Hu, X. He, J. Gao, Ms-celeb-1m: A dataset and benchmark for large-scale face recognition, in *Computer Vision – ECCV 2016*, Springer, (2016), 87–102. https://doi.org/arXiv:1607.08221

41. Y. Li, Y. Lu, J. Li, G. Lu, Separate loss for basic and compound facial expression recognition in the wild, in *Proceedings of The Eleventh Asian Conference on Machine Learning*, (2019), 897–911.

42. K. Wang, X. Peng, J. Yang, S. Lu, Y. Qiao, Suppressing uncertainties for large-scale facial expression recognition, *arXiv preprint*, (2020), arXiv:2002.10392. https://doi.org/10.48550/arXiv.2002.10392

43. Z. Zhao, Q. Liu, F. Zhou, Robust lightweight facial expression recognition network with label distribution training, in *Proceedings of the AAAI conference on artificial intelligence*, **35** (2021), 3510–3519. https://doi.org/10.1609/aaai.v35i4.16465

44. E. Barsoum, C. Zhang, C. C. Ferrer, Z. Zhang, Training deep networks for facial expression recognition with crowd-sourced label distribution, in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, (2016), 279–283. https://doi.org/10.1145/2993148.2993165

45. C. Huang, Combining convolutional neural networks for emotion recognition, in *2017 IEEE MIT Undergraduate Research Technology Conference (URTC)*, (2017), 1–4. https://doi.org/10.1109/URTC.2017.8284175

46. S. Albanie, A. Nagrani, A. Vedaldi, A. Zisserman, Emotion recognition in speech using cross-modal transfer in the wild, *arXiv preprint*, (2018), arXiv:1808.05561. https://doi.org/10.48550/arXiv.1808.05561

AIMS Press