Applied Intelligence (2022) 52:2918–2929
https://doi.org/10.1007/s10489-021-02575-0

A new multi-feature fusion based convolutional neural network for facial expression recognition

Wei Zou[1] · **Dong Zhang**[1] · **Dah-Jye Lee**[2]

Accepted: 17 April 2021 / Published online: 25 June 2021

Abstract

Using lightweight networks for facial expression recognition (FER) is becoming an important research topic in recent years. The key to the success of FER with lightweight networks is to explore the potentials of expression features in distinct abstract levels and regions, and design robust features to characterize the facial appearance. This paper proposes a lightweight network called Multi-feature Fusion Based Convolutional Neural Network (MFF-CNN), for image-based FER. The proposed model uses the Image Branch to extract both mid-level and high-level global features from the whole input image and utilizes the Patch Branch to extract local features from sixteen image patches of the original image. In MFF-CNN, feature selection based on L2 norm is performed to obtain more discriminative local features. Joint tuning is employed to integrate the two branches and fuse features. Experiment results on three widely used datasets, CK+, JAFFE and Oulu-CASIA show the proposed MFF-CNN outperforms the state-of-the-art methods in terms of average recognition accuracy. Compared to other competitive models with similar or larger number of parameters, our MFF-CNN improves the average recognition accuracy by 9.80% to 15.05%.

Keywords Facial expression recognition · Multi-feature fusion convolutional neural network · Feature selection · Joint tuning

1 Introduction

Facial expression recognition (FER) has been a popular research topic in computer vision for decades. Various systems based on facial expression recognition, including support system for diagnosis of neurological disorders [1], driver-assistance system [2] and face-to-face neural conversation model [3], have already been applied in our daily life.

✉ Dong Zhang
zhangd@mail.sysu.edu.cn

Wei Zou
zouw23@mail2.sysu.edu.cn

Dah-Jye Lee
djlee@byu.edu

[1] School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou, Guangdong Province, 510006, China

[2] Department of Electrical and Computer Engineering, Brigham Young University, Provo, Utah, 84602, USA

The studies on facial expression analysis can be dated back to Ekman and Friesen's work [4], in which human facial expressions were grouped into six basic categories, including anger, disgust, fear, happiness, sadness and surprise. General approaches of facial expression recognition are usually categorized into two groups: Action Unit (AU) based and image feature based. In the Facial Action Coding System (FACS) [5], AUs refer to a couple of muscle movements around facial organs. Combinations of AUs are used to account for all possible facial expressions. AU based methods attempt to detect and use specific descriptors from the facial images for facial expression recognition [6–8]. Image feature-based methods represent facial expression image with certain patterns that characterizing the hidden information around facial organs or landmark points [9–11].

Among the traditional image feature-based FER methods, many effective hand-crafted feature descriptors were used to characterize the expression image, e.g. Local Binary Patterns (LBP) [12], Local Description Patterns (LDP) [13], and Scale Invariant Feature Transform (SIFT) [14]. However, variations including illumination, rotations, and noise in images may weaken the capacity of such hand-crafted

feature descriptors and thus affect the performance of the traditional FER approaches.

In the last decade, deep learning-based approaches have achieved great success in many recognition tasks using image, speech and video. Some large-scale deep neural networks [15, 16] also obtained impressive accuracy in facial expression recognition owing to their strong feature extraction ability. However, the limitation of available samples in facial expression recognition datasets makes the training of these large-scale deep neural networks prone to over-fitting. Furthermore, the implementation of large-scale deep learning networks requires a large amount of computational resources and highly standard configuration. This also restricts its application for resource-limited platforms.

In recent years, many lightweight networks with fewer layers were proposed for the task of FER. In order to improve recognition performance, these networks explored facial expression features from various perspectives. Some works regarded the expression as an integrated motion of muscles in human face, and extracted features of facial expression from the whole face image [20, 21]. While other works focused more on local features of human face [17, 34]. Xie et al. proposed a two-branch networks to extract local and global features from image patches and the whole image respectively for FER [17]. The combination of local and global features improved the performance and obtained promising average accuracy on facial expression recognition. From another perspective, many research works tried to explore facial expression features from distinct abstract levels. Some networks only utilized high-level features generated from the last layer of the network to make final predictions [28, 30], while others combined high-level features with mid-level features extracted by intermediate layers [18, 19]. Nguyen et al. designed a multi-level Convolutional Neural Networks (CNNs) that extracts both mid-level and high-level features for the classification of facial expressions [18]. The involvement of mid-level features enhances the performance of networks in recognizing facial expression in the wild. Furthermore, based on the visualization result of the feature maps, they concluded that mid-level features are capable of representing basic semantic abstract of original inputs and are robust against noise [18].

Literature review shows that the key to the success of FER is to explore the potentials of expression image in distinct abstract levels and regions, and design robust feature to characterize the facial appearance. This paper proposes a lightweight network, named Multi-feature fusion based Convolutional Neural Network (MFF-CNN), to utilize different efficient features proposed in state-of-the-art methods for FER in constrained environments. As shown in Fig. 1, our proposed MFF-CNN extracts facial expression features through two parallel branches, Image Branch and Patch Branch. The Image Branch captures global features in the mid and high levels from the whole input image, while the Patch Branch segments the input image into sixteen overlapping image patches and extracts local features from each of them. Human facial expression is categorized according to the fused features from the Image Branch and Patch Branch. To evaluate the performance of MFF-CNN for facial expression recognition, we carried out experiments on three widely used facial expression datasets: CK+, JAFFE and Oulu-CASIA. On these three datasets, the average recognition accuracy of our proposed MFF-CNN is higher than other state-of-the-art methods.

The contribution of this paper is summarized as follow:

1. A two-branch model called MFF-CNN is proposed to improve the accuracy of facial expression recognition by fusing diverse features from facial expression images. Image Branch extracts mid-level and high-level global features from the whole image, while Patch Branch extracts local features from overlapping image patches. In addition, joint tuning is utilized to integrate these two branches, and improve the recognition accuracy.
2. We employ L2-norm based feature selection to obtain more discriminative local features and reduce the number of parameters in the Patch Branch.

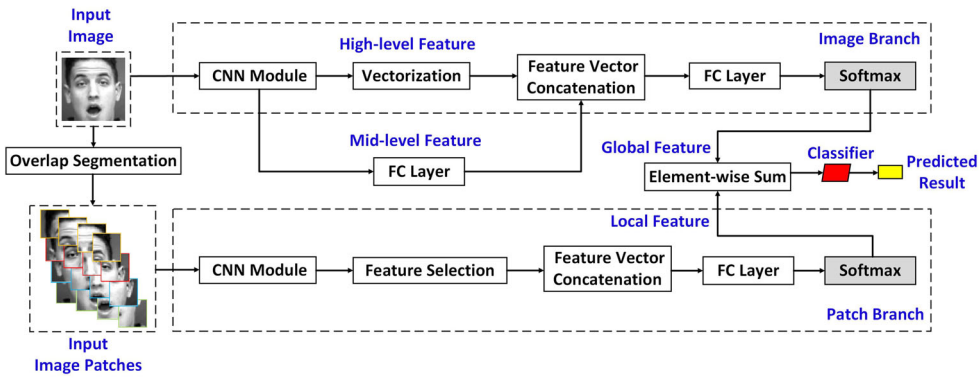

Fig. 1 Framework of our proposed method

3. The proposed MFF-CNN model obtained better recognition accuracy than the state-of-the-art models on three popular facial expression datasets. The improvement on the average recognition accuracy is from 9.80% to 15.05% on the Oulu-CASIA datasets, compared to other competitive models with similar or larger number of parameters.

The rest of the paper is structured as follow: Section 2 introduces related work on facial expression recognition briefly. The details of our proposed framework are presented in Section 3. In Section 4, experimental results are discussed and compared with other related work. At last, we conclude our work in Section 5.

2 Related work

In the field of computer vision, it is a big challenge to correctly recognize the expression from facial image. Traditional hand-crafted feature-based approaches for FER depend heavily on researcher's experience and their performance is strongly affected by image quality variations.

Deep learning-based methods have been applied in the field of facial expression recognition to solve this problem. Hamester et al. used a Multi-channel CNN (MCCNN) to recognize facial expressions [20]. In MCCNN, both CNN channel and convolutional autoencoder (CAE) channel shared the same topology. The first convolutional kernel in the CAE channel is pre-trained and the parameters are kept fixed during the model training stage. The recognition accuracy of MCCNN outperformed the traditional hand-crafted feature-based methods.

Although large scale deep learning models obtain good recognition accuracy, the limited number of samples in facial expression datasets constraints their performance. Without adequate training samples, large scale deep learning models are prone to over-fitting. To obtain a balance between architecture complexity and recognition accuracy, researchers try to design lightweight deep learning models with a compact structure and strong ability to extract features. In order to obtain comparable or superior performance to large scale networks, lightweight models are required to explore potential expression image in distinct regions, and construct multi-scale features to characterize facial appearance.

Xie et al. proposed a two-branch CNN framework, Deep Comprehensive Multi-Patch Aggregation Convolutional Neural Network (DCMA-CNN) for FER [17]. One branch of DCMA-CNN extracts holistic features from complete facial expression images, while the other branch is used to extract local features from a group of overlapping patches. After feature extraction, the holistic features and local features are concatenated together for final prediction. As the aggregation of both local and holistic features represents facial expressions at different scales, the recognition accuracy of DCMA-CNN is better than other competitive methods for facial expression classification.

Besides exploring hidden information in distinct regions, other lightweight models have also committed to extracting discriminative features in diverse abstract levels from the facial expression image. Nguyen et al. presented a multi-level CNN (MLCNN) model for FER [18]. Previous work [19] found that besides the high-level features generated by the last convolutional layers, low-level and mid-level features extracted by the primary and intermediate convolutional layers also provide useful information for FER. They bypassed low-level features because low-level features mostly represent the elementary textures from the images and focus more on the background and other meaningless regions. While mid-level features can capture basic abstract of input images and remain discriminative under the interference of noise. Thus, in MLCNN, two fully connected layers were inserted to utilize both mid-level and high-level features for classification. The combination of mid-level and high-level features improved the robustness of the model against variations on face pose and image quality and obtained high recognition accuracy in recognizing facial expression in the wild.

In addition to feature extraction, model integration is another way to improve the performance of lightweight deep learning models on FER. To aggregate features extracted by different parts of the model, most proposed methods formed the features for final recognition by feature vectors concatenation [17, 18, 20]. Although these methods are simple to implement, different parts of the integrated model may focus on their distinct regions of interest and obtain incorrect predictions. Jung et al. proposed to use joint fine-tuning to integrate features extracted from different parts of model [21]. They pre-trained two deep networks and froze the weights in each network except the top fully connected layers. In the training stage, the top fully connected layers were retrained and fine-tuned with back propagation based on redefined loss function. The final prediction was made based on the element-wise sum of the two networks. This new method boosted the performance of their proposed model on facial expression recognition, compared to other state-of-the-art models.

Literature review shows that features extracted from mid-level and high-level layers can provide robust and discriminative description of the facial expression, while features captured from local patches may form a learnable attention model and contribute to the improvement of the network performance. This observation motivates us to fuse efficient features discovered by state-of-the-art models and

maximize their contributions to improve facial expression recognition accuracy.

3 Proposed method

In this paper, we propose a new network named MFF-CNN to integrate efficient features discovered by state-of-the-art works for FER. The proposed network includes an Image Branch and a Patch Branch. The two branches are integrated by joint tuning.

The architecture of MFF-CNN is shown in Fig. 2. The Image Branch of the network characterizes the input image with mid-level and high-level features from a global perspective. While the Patch Branch extracts local feature from a number of overlapping image patches. Feature vectors selection is also employed to obtain more discriminative local features and decrease the compute nodes in subsequent fully connected layers. The feature vector selection mechanism significantly reduces the number of parameters in the model. At last, the Image Branch and the Patch Branch are integrated with joint tuning to properly fuse features generated by the two branches as well as improve the accuracy of expression recognition. The proposed MFF-CNN is a lightweight network as the number of parameters is 1.21M and the number of floating- point operations (FLOPs) is 4.87G.

3.1 Image branch

As shown in Fig. 2, the CNN module in the Image Branch of our proposed MFF-CNN contains six convolutional layers and three max-pooling layers to extract global features from the original images. Each of the convolutional layer utilizes ReLU as the activation function.

Some research works have discovered that mid-level features extracted by intermediate convolutional layers provide basic semantic abstract of original inputs and are discriminative under noise interference [18, 19]. Thus, in the Image Branch of our proposed MFF-CNN model,

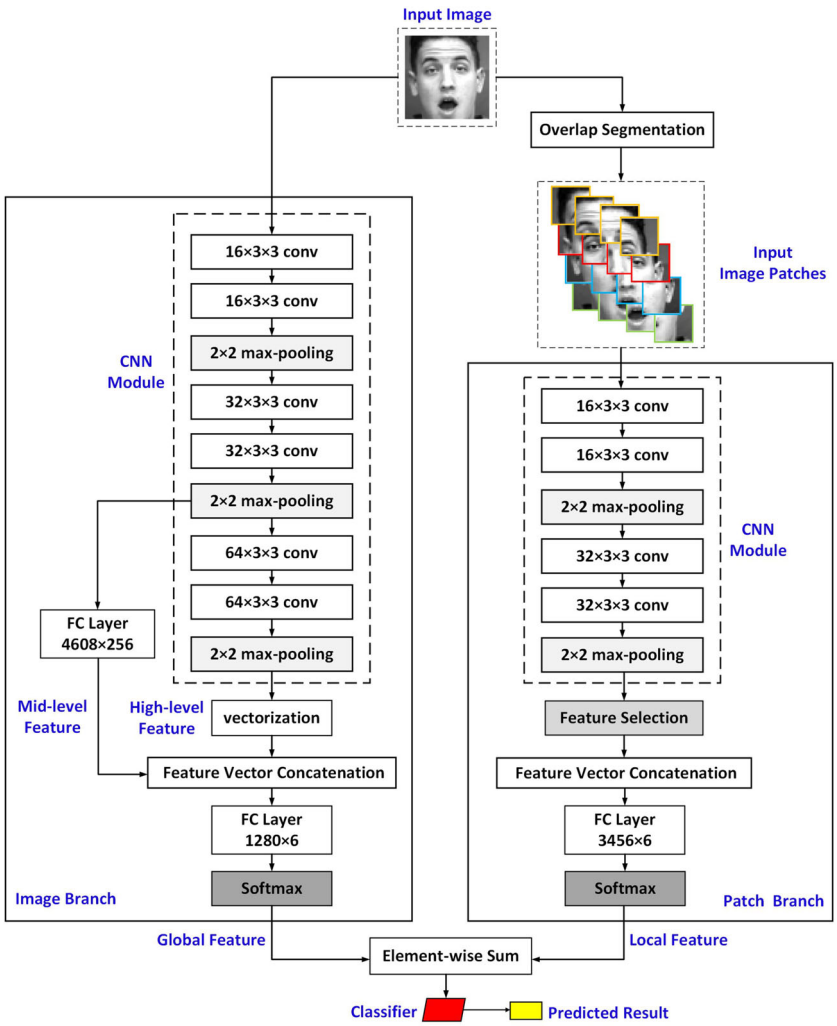

Fig. 2 The architecture of MFF-CNN

we employ both high-level features generated by the last convolutional layer and mid-level features extracted by the 4^{th} convolutional layer. An extra fully-connected layer with 256 units is employed to reserve efficient mid-level features. Vectorization is also performed to change output high-level feature maps into feature vectors. The feature vector concatenation operation which combines the mid-level features and the high-level features is defined as (1):

$$(x_1^{mid}, x_2^{mid}, ...)^T \bigoplus (x_1^{high}, x_2^{high}, ...)^T = (x_1^{mid}, x_2^{mid}, ..., x_1^{high}, x_2^{high}, ...)^T \quad (1)$$

where "$\bigoplus$" refers to feature vector concatenation operator, $(x_1^{mid}, x_2^{mid}, ...)^T$ refers to the mid-level feature and $(x_1^{high}, x_2^{high}, ...)^T$ refers to the high-level feature.

Then, the concatenated global feature is fed into a fully-connected layer and a softmax classifier to generate the prediction of the possibility that the input image belongs to each kinds of facial expressions.

3.2 Patch branch

Many previous studies [9, 11, 17, 22] have shown the effectiveness of features generated from certain local areas of facial image. These discriminative features are regarded as the complementary information to the global feature captured from the whole image. A local feature can characterize the specificity of local region, while a group of several local features will provide a pattern of attention for the whole facial expression image.

To extract local features of input images, we first segment the input images into a series of overlapping square patches. The overlap ratio between two neighboring patches is set to 0.5. Fig. 3 displays the effect of our overlap segmentation operation with one sample from the CK+ dataset. In order to ensure the correctness of this overlap segmentation operation, the edge length of each patch and original image must satisfy (2):

$$S \times (\sqrt{N} + 1) = 2L \quad (2)$$

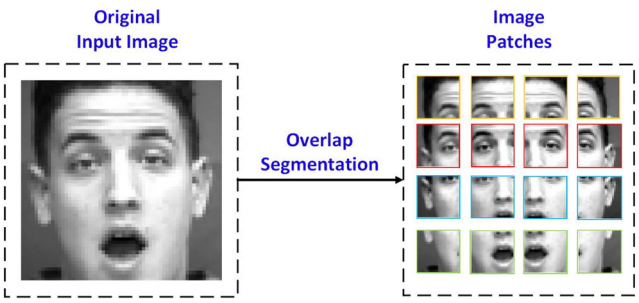

Fig. 3 The effect of overlap segmentation

where S is the edge length of each patch, L is the edge length of original image and N is the number of patches. N is a perfect square number. In this paper, N is set to 16.

After overlap segmentation, the CNN module in the Patch Branch is used to capture local features from these N patches one by one. As shown in Fig. 2, the structure of the CNN we use in the Patch Branch is similar to that used in Image Branch, but includes four convolutional layers and two max-pooling layers.

As the CNN module in the Patch Branch extracts a large number of features and leads to a large number of parameters in the following fully-connect layer, we employ a mechanism of feature selection to obtain more discriminative local features and decrease the compute nodes in the following fully connected layers. In [23], the authors compared the performance of aggregating features with largest norm or aggregating the same amount of randomly selected features for image classification. They concluded that deep features with large norm are more discriminative than features with random norm [23]. Experiment results in [17] also showed that selecting feature vectors with higher magnitude of L2-norm and bypassing feature vectors with small norm can have a better performance on facial expression recognition.

In this work, the most discriminative local features are selected using their L2-norm. The general procedure is summarized in three steps. First, we vectorize the output feature maps from the CNN module in the Patch Branch, and obtain $N \times K$ feature vectors. N refers to the number of image patches and K refers to the number of feature maps extracted from one patch. In this paper, N is set to 16 and K is set to 32, and the total number of features extracted by the CNN module in the Patch Branch is 512. Second, we evaluate the L2-norm of all feature vectors. Last, we select M feature vectors with the largest L2-norm. We set M to 384 in this paper.

After feature selection, we perform feature vector concatenation on the selected local feature vectors. The operation is the same as (1). Then, like the Image Branch, the concatenated local feature is fed into a fully-connected layer and a softmax classifier to generate the prediction of the Patch Branch.

3.3 Integration

Although feature vector concatenation is widely used in many multi-channel or multi-branch architectures [17, 18, 20], it is not suitable for the proposed MFF-CNN. As the Image Branch and the Patch Branch focus on different regions of interest, the concatenation of features from the Image Branch and Patch Branch may result in diverse prediction.

We perform joint tuning to integrate the extracted features of the Image Branch and Patch Branch in our proposed MFF-CNN. First, we carry out element-wise sum operation to the output of the two branches as (3):

$$\tilde{y_i} = l_{image,i} + l_{patch,i} \tag{3}$$

where $l_{image,i}$ is the prediction of the Image Branch that the input images belongs to the i^{th} facial expressions, $l_{patch,i}$ is the prediction of the Patch Branch, $\tilde{y_i}$ is the integrated possibility that the input image belongs to the i^{th} facial expressions. Then, the final prediction of our framework can be formulated by (4):

$$o = \arg\max_i \tilde{y_i} \tag{4}$$

In training stage, the parameters of our proposed MFF-CNN framework are randomly initialized. Then, error back propagation is applied to tune all the parameters. The loss function we used is defined as (5):

$$L_{MFF-CNN} = \lambda_1 \left(-\sum_i t_i \log l_{image,i} \right) + \lambda_2 \left(-\sum_i t_i \log l_{patch,i} \right) + \lambda_3 \left(-\sum_i t_i \log \tilde{y_i} \right) \tag{5}$$

where t_i is the truth label of each image, i.e. the probability belonging to the i^{th} facial expression. In this paper, the label of each sample is coded in one-hot form. The value of t_i can be 0 or 1. λ_1, λ_2 and λ_3 are tuned parameters, and set to 1, 1, 0.1 in our experiments.

4 Experiments

To evaluate the performance of our proposed MFF-CNN model, we carried out experiments on three datasets collected in constrained environments: CK+ [24], JAFFE [25], and Oulu-CASIA [26], and one dataset collected in-the-wild, SFEW2.0 [27]. These three datasets are publicly available and widely used in research of facial expression recognition. In order to compare with other existing methods, we chose images with one of six basic classes of expressions, including anger, disgust, fear, happy, sadness, and surprise, from each dataset.

4.1 Datasets and preprocessing

The CK+ dataset includes 529 image sequences captured from 123 subjects and 327 of them are annotated with eight expression labels. All image sequences start with a neural expression and gradually transit into peak expression. Following the experiment routines [9, 28, 29], we selected the last three frames with peak expression in sequences with six basic expression labels. Thus, 309 image sequences were included in our experiments and we obtained a total of 927 samples from this dataset. The distribution of selected samples of the three datasets we used in our experiments is shown in Table 1.

The JAFFE dataset contains 213 images taken from 10 Japanese females. 30 of the images are labeled with neutral expression. In our experiment, 183 images labeled with six basic expressions were chosen to compare with other state-of-the-art methods. Due to the relatively small number of samples, we also applied data augmentation to avoid over-fitting. We flipped the original images horizontally and rotated them by angles of 5, 10 and 15 degrees clockwise and counterclockwise. With data augmentation, we obtained seven diverse versions of samples besides original images, and the total number of samples from this dataset for our experiments was 1464.

The Oulu-CASIA dataset has total 2880 image sequences taken from 80 subjects under three illumination conditions (normal, weak and dark) and cameras in visible and near-infrared ranges. Each sequence is labeled with one of the six basic expressions. Following the experiment routines of [29] and [30], we selected 480 image sequences captured by VIS camera under normal indoor illustration condition and the last three frames were used in experiment. At last, 1440 samples from the Oulu-CASIA dataset were included in our experiment.

Furthermore, to reduce the impact of background, we applied Viola-Jones face detector [31] to detect human face in each sample. The facial region was resized to 60×60. Then, the facial expression images were randomly divided into ten subsets with equal size for the evaluation of 10-fold cross-validation.

Table 1 The number of samples selected in our experiments for each facial expression

Dataset	Anger	Disgust	Fear	Happy	Sadness	Surprise	Total
CK+	135	177	75	207	84	249	927
JAFFE	240	232	256	248	248	240	1464
Oulu-CASIA	240	240	240	240	240	240	1440

4.2 Experiment settings

To evaluate the performance of our proposed MFF-CNN framework, we conducted a series of experiments with Pytorch platform. Cuda 8.0 was also employed to speed up model training.

In the first stage, the parameters of our model were randomly initialized, and the initial learning rate was set to 0.01. Adam [33] was used as the optimization algorithm, and the batch size was set to 100. The training epochs for CK+, JAFFE and Oulu-CASIA dataset were 25, 40 and 40, respectively. During the model training stage, learning rate decayed in an exponential form as shown in (6):

$$\eta = \eta_0 \times 0.93^{\lfloor \frac{i}{3} \rfloor} \tag{6}$$

where η_0 is the initial learning rate, i refers to the iterative epoch of training stage and symbol $\lfloor . \rfloor$ stands for rounding down operation.

4.3 Experiment on CK+ dataset

The experiment results of our proposed MFF-CNN on CK+ dataset are given in Fig. 4 and Table 2. Figure 4 shows the prediction results of different expressions in a confusion matrix. The comparison of the average recognition accuracy with other state-of-the-art models is listed in Table 2.

Figure 4 shows that our MFF-CNN model performed better in recognizing disgust, happy and surprise, with the average recognition accuracy over 99%. However, the average accuracies of anger and sadness were relatively lower than other expressions. Some samples of these two expressions were easily misclassified by each other. Further investigation on the misclassified image samples shows that these two expressions share some similar patterns such as the wrinkling of nose and the lowering of the upper lip. As shown in Facial Action Coding System, anger relates to six kinds of AUs and three of them are identical to sadness [5].

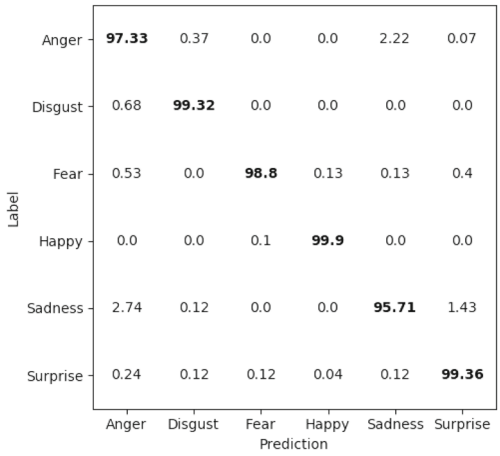

Fig. 4 Confusion matrix of MFF-CNN on CK+ dataset. The vertical axis is the ground truth label and the horizontal axis is the prediction

Table 2 Comparison of accuracy and model size with existing models on CK+ dataset

Model	Parameters	Accuracy(%)
DCMA-CNN [17]	0.05M	93.46
DTAGN [21]	5.85M	97.25
FN2EN [28]	1.19M	98.60
MFF-CNN	1.21M	98.80

Table 2 shows that our proposed MFF-CNN model obtained an average recognition accuracy of 98.80% on CK+ dataset, which was higher than other existing state-of-the-art models. The number of parameters for MFF-CNN is much larger than DCMA-CNN but it imptoves the accuracy by more than 5%. This is because MFF-CNN inserts an extra fully connected layer in the Image Branch, which requires a large amount of parameters. It takes mid-level and high-level features together into consideration, to obtain performance improvement. Compared to DTAGN, MFF-CNN reduces the model size by about 80% but still achieves about 1.5% improvement in accuracy. Finally, the number of parameters for MFF-CNN is similar to FN2EN, but the recognition accuracy is improved by 0.2%. The reason is that FN2EN only utilizes complete images to extract features but ignores discriminative local area.

4.4 Experiment on JAFFE dataset

Similar to the evaluation on CK+ dataset, we depict the confusion matrix of each expression on JAFFE dataset in Fig. 5. The result from comparison with other competitive models is given in Table 3.

Figure 5 shows the proposed MFF-CNN model obtained impressive recognizing accuracy for anger (98.83%), happy

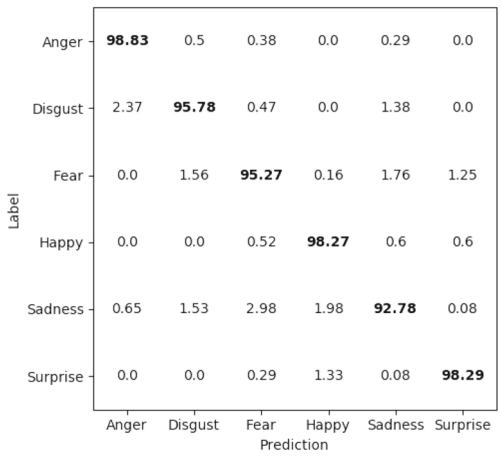

Fig. 5 Confusion matrix of MFF-CNN on JAFFE dataset

Table 3 Comparison of accuracy and model size with existing models on JAFFE dataset

Model	Parameters	Accuracy(%)
Sobel-based CNN [20]	0.58M	92.00
CAE-based CNN [20]	0.43M	94.10
DCMA-CNN [17]	0.05M	94.75
MFF-CNN	1.21M	96.52

Table 4 Comparison of accuracy and model size with existing models on Oulu-CASIA dataset

Model	Parameters	Accuracy(%)
DTAGN [21]	5.85M	81.46
FaceLiveNet [30]	1.31M	87.50
FN2EN [28]	1.19M	87.71
MFF-CNN	1.21M	96.51

(98.27%) and surprise (98.29%). While, the lowest average recognition accuracy was for sadness with 92.78%. Nearly 3% of the samples for sadness were confused with fear. The reason is the samples of sadness and fear on JAFFE dataset share certain similar appearance in several local facial regions, e.g. inner eyebrows, upper lip, and area around the nose.

We compared the accuracy and the model size of our proposed MFF-CNN with other state-of-the-art methods in Table 3. Table 3 shows the model size, in terms of number of parameters, of the proposed MFF-CNN was larger than other state-of-the-art models. However, MFF-CNN obtained the highest performance in terms of the average recognition accuracy with the improvement ranges from 1.77% to 4.52%. Compared with these competitive methods, the proposed MFF-CNN involves mid-level features for classification. Although this technique employs a fully-connected layer with around 1.18M parameters, it utilizes additional discriminative global features and improves the accuracy of FER.

4.5 Experiment on Oulu-CASIA dataset

On the Oulu-CASIA dataset, we carried out a similar experiment as mentioned in Sections 4.3 and 4.4. Figure 6 displays the confusion matrix of different expressions and Table 4 lists the result from comparison with other state-of-the-art models.

As shown in Fig. 6, our proposed MFF-CNN model performed the best in recognizing fear, happy and surprise, with the average recognition accuracy over 98.5%. While the recognition accuracies for anger, disgust, and sadness were lower than 95%. Figure 6 also shows over 4% of the disgust and sadness expression samples were categorized into anger. Further investigation on these misclassified image samples showed that similar muscle movement including wrinkling around nose and upper lip, contraction between eyebrows appear in these three expressions.

Table 4 shows that our proposed MFF-CNN model outperformed all state-of-the-art models in terms of recognition accuracy on the Oulu-CASIA database. The average recognition accuracy of MFF-CNN ranges from 9.80% to 15.05% higher than the existing competitive methods with similar or slightly more parameters. Even though the MFF-CNN has around 0.02M more parameters than FN2EN, the improvement in recognition accuracy demonstrates its effectiveness.

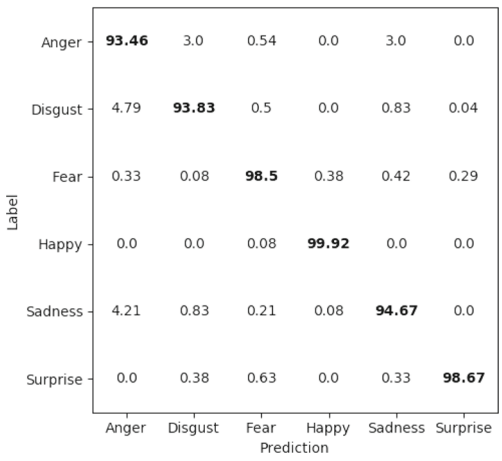

Fig. 6 Confusion matrix of MFF-CNN on Oulu-CASIA dataset

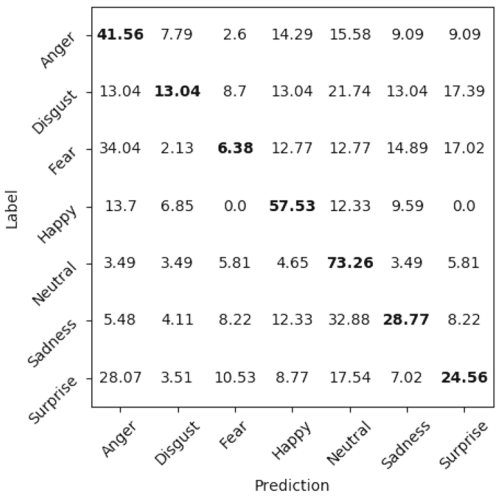

Fig. 7 Confusion matrix of MFF-CNN on SFEW2.0 dataset

Table 5 Comparison of accuracy and model size with models specifically designed for unconstrained facial expression recognition on SFEW2.0 dataset

Model	Parameters	Accuracy(%)
DLP-CNN [35]	73.75M	51.05
IBAN-IL [36]	19.8M	55.28
MFF-CNN	1.21M	40.83

4.6 Experiment on SFEW2.0 dataset

Although the proposed method is designed for FER in constrained environments, we evaluated its performance on an in-the-wild dataset, SFEW2.0. Different from the CK+, JAFFE and Oulu-CASIA datasets, SFEW2.0 was created for unconstrained facial expression recognition. The samples in SFEW2.0 were extracted from the key frames of video clips and label with seven expressions (anger, disgust, fear, happy, neutral, sadness and surprise). SFEW2.0 dataset was divided into training, validation and test sets.

For images in SFEW 2.0, we used Retina-Face [32] to detect human face, and the facial region was also resized to 60×60. After preprocessing, we first used all the samples collected from the CK+, JAFFE and Oulu-CASIA datasets to pretrain our model. The training epoch was set to 30. Then, we trained our model with the samples from the training set of the SFEW2.0 dataset for 200 epochs. Other experiment settings remained the same as those in Section 4.2.

Following the experiment procedures of [35] and [36], we evaluated the performance of our proposed model on the validation set. Figure 7 shows the confusion matrix of seven expressions and Table 5 lists the comparison with other state-of-the-art models designed for unconstrained facial expression recognition.

Figure 7 shows our proposed MFF-CNN obtained adequate performance of recognizing neutral (73.26%), happy (57.53%) and anger (41.56%). The recognition accuracies for fear and disgust are much lower and less than 15%.

Table 5 shows that the recognition accuracy obtained by MFF-CNN is not as good as those state-of-the-art models [35, 36] which were designed specifically for unconstrained facial expression recognition. In our MFF-CNN, we prefer to use as few convolutional layers as possible to construct an efficient and lightweight network, which limits its performance when it deals with unconstrained FER task. Our proposed model was designed to operate in constrained environments.

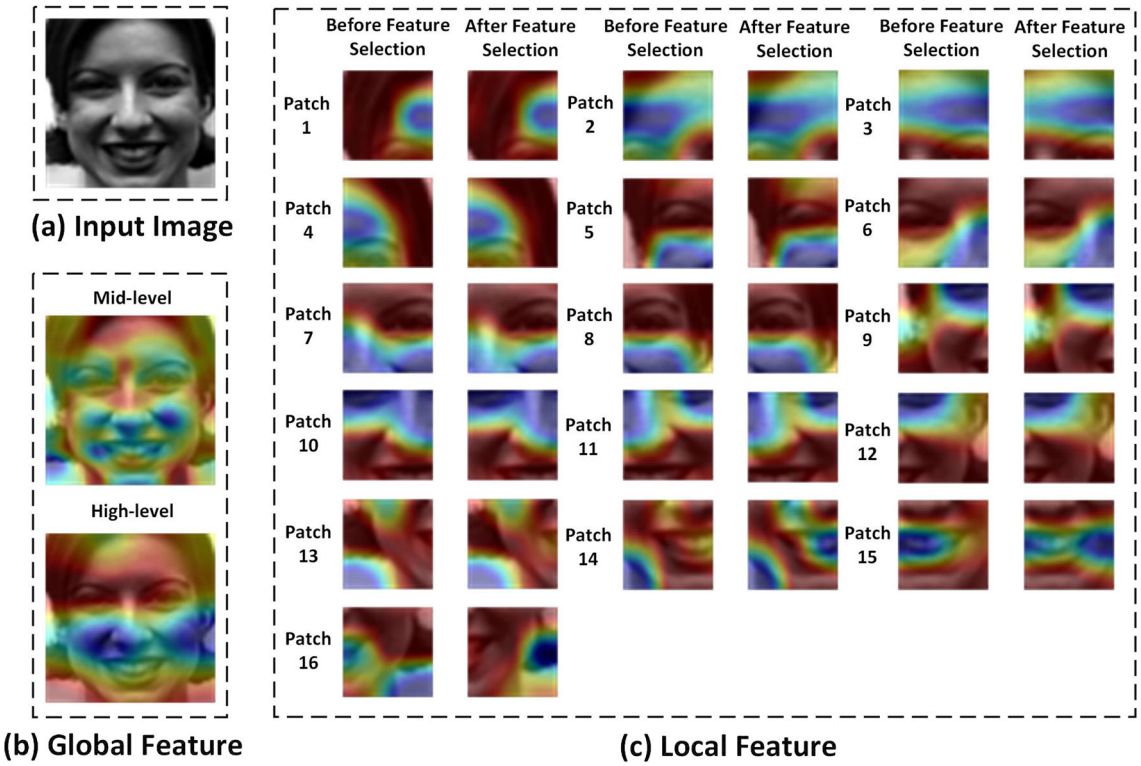

Fig. 8 Grad-CAM visualization results of different kinds of feature maps generated by MFF-CNN

Table 6 The number of samples selected in our experiments for each facial expression

Method	Parameters	Ang	Dis	Fea	Hap	Sad	Sur	Avg
MFF-CNN	1.21M	93.46	93.83	98.50	99.92	94.67	98.67	96.51
MFF-CNN without mid-level feature	0.02M	70.69	83.89	86.39	96.53	82.22	96.25	86.00
MFF-CNN without joint tuning	1.21M	92.17	93.54	98.42	99.96	94.70	98.88	96.26
MFF-CNN without feature vector selection	1.22M	91.04	92.67	98.46	99.83	94.67	98.88	95.92

4.7 Visualization

In order to explore how global features and local features represent the input image, we applied Gradient-weighted Class Activation Mapping (Grad-CAM) [37] to visualize the region of interest in different feature maps. Grad-CAM computes the gradients of target result with respect to the selected feature maps, and obtains the weights indicating the importance of feature maps for target result. Then, the weights are multiplied to the feature map to obtain a coarse localization map which highlights the important regions for predicting the result [37].

In the Image Branch, we selected feature maps extracted by the 2^{nd} max-pooling and 3^{rd} max-pooling layer as mid-level and high-level global features, respectively for Grad-CAM. In the Patch Branch, we used both feature maps generated by the 2^{nd} max-pooling layer and feature maps received from feature selection, for Grad-CAM to visualize the effect of our L2-norm based feature selection method. Then, we multiplied the Grad-CAM results to the original input image for visualization. Figure 8 displays the visualization results of one random sample from CK+ dataset.

In each Grad-CAM visualization image in Fig. 8, the warmer color represents higher attention level and the colder color represents lower attention level. For visualization of global feature, Fig. 8(b) shows that for mid-level global features (top), regions with high attention levels are more separated. By contrast, for high-level global features (bottom), regions with high attention levels are more concentrated. These regions distribute over forehead, bridge of the nose and jaw. For visualization of local features, Fig. 8(c) shows distinct patches focus on different local regions of the input expression image. In Patch 14, before feature selection, mouth and nose are regarded as regions with high attention levels. After feature selection, same areas change into regions with low attention levels.

4.8 Ablation research

To evaluate the contribution of mid-level feature aggregation, joint tuning, and feature vector selection, we carried out a series of ablation experiments based on the Oulu-CASIA dataset. As shown in Table 6, in MFF-CNN without mid-level features, we only used high-level features to represent global features. In MFF-CNN without joint tuning, we integrated the two branches with traditional feature vector concatenation. In MFF-CNN without feature vector selection, we used complete local feature vector set to represent local features.

Table 6 shows any deletion of mid-level feature aggregation, joint tuning or feature vector selection, led to the decrease of the average recognition accuracy of our MFF-CNN. Table 6 also shows the using of mid-level feature aggregation improved the average recognition accuracy from 86.00% to 96.51%. For the most challenging expression classification tasks in the Oulu-CASIA dataset, i.e. anger, disgust and sadness, mid-level feature aggregation helped the recognizing accuracy of these three expression remarkably, from 70.69% to 92.46% for anger, from 83.89% to 93.83% for disgust, and from 82.22% to 94.67% for sad. The great improvements of recognition accuracy demonstrated that mid-level features aggregation made significant contributions to FER.

5 Conclusion

In this paper, we propose a two-branches model named MFF-CNN for facial expression recognition in constrained environments. Our proposed model MFF-CNN is a small-scale model with fewer layers than other competitive deep learning models. It fuses the discriminative features proposed in state-of-the-art methods and obtains high recognition accuracy.

In MFF-CNN, the Image Branch captures both mid-level and high-level features to characterize facial expression from a global perspective. The Patch Branch extracts local features from sixteen overlapping patches to provides complimentary description of the image patches. Rather than using the traditional feature vector concatenation, we utilize joint tuning to integrate the features generated from Image Branch and Patch Branch. To reduce the number of parameters ofthe proposed model, feature selection based on the L2-norm of each feature vector is employed.

Compared with the state-of-the-art networks, experiment results show that for the CK+ and JAFFE datasets, the proposed network obtains improved recognition accuracy

with comparable model size. Particularly, for the Oulu-CASIA dataset, our proposed model obtained 9.80% to 15.05% improvement on recognition accuracy, compared to other competitive models with slightly more or comparable number of parameters.

Acknowledgements This work was supported by Guangzhou Municipal People's Livelihood Science and Technology Plan (201903010040), and Science and Technology Program of Guangzhou, China (202007030011).

References

1. Yolcu G, Oztel I, Kazan S et al (2019) Facial expression recognition for monitoring neurological disorders based on convolutional neural network. Multimed Tools Appl 78:31581–31603. https://doi.org/10.1007/s11042-019-07959-6
2. Jabon M, Bailenson J, Pontikakis E et al (2011) Facial expression analysis for predicting unsafe driving behavior. IEEE Perv Comput 10:84–95. https://doi.org/10.1109/mprv.2010.46
3. Chu H, Li D, Fidler S (2018) A face-to-face neural conversation model. In: IEEE/CVF Conference on computer vision and pattern recognition(CVPR), pp 7113–7121. https://doi.org/10.1109/cvpr.2018.00743
4. Ekman P, Friesen WV (1971) Constants across cultures in the face and emotion. J Pers Soc Psychol 17(2):124–129. https://doi.org/10.1037/h0030377
5. Ekman P, Friesen WV (1978) Facial action coding system (FACS): A technique for the measurement of facial movement. Consulting Psychologists Press
6. Wang S, Ding H, Peng G (2020) Dual learning for facial action unit detection under nonfull annotation. IEEE Trans Cybern 1–13. https://doi.org/10.1109/TCYB.2020.3003502
7. He J, Yu X, Sun B, Yu L (2021) Facial expression and action unit recognition augmented by their dependencies on graph convolutional networks. J Multimodal User Interfaces. https://doi.org/10.1007/s12193-020-00363-7
8. Wang S, Peng G (2019) Weakly supervised dual learning for facial action unit recognition. IEEE Trans Multimed 21:3218–3230. https://doi.org/10.1109/TMM.2019.2916063
9. Zhong L, Liu Q, Yang P et al (2007) Learning Multiscale Active Facial Patches for Expression Analysis. IEEE Trans Cybern 45:1499–1510. https://doi.org/10.1109/tcyb.2014.2354351
10. Majumder A, Behera L, Subramanian VK (2018) Automatic facial expression recognition system using deep network-based data fusion. IEEE Trans Cybern 48:103–114. https://doi.org/10.1109/tcyb.2016.2625419
11. Majumder A, Behera L, Subramanian VK (2018) Emotion recognition from geometric facial features using self-organizing map. Pattern Recognit 47:1282–1293. https://doi.org/10.1016/j.patcog.2013.10.010
12. Kong F (2019) Facial expression recognition method based on deep convolutional neural network combined with improved LBP features. Pers Ubiquitous Comput 531–539. https://doi.org/10.1007/s00779-019-01238-9
13. Revina IM, Emmanuel WRS (2019) Face expression recognition with the optimization based multi-SVNN classifier and the modified LDP features. J Vis Communi Image Represent 62:43–55. https://doi.org/10.1016/j.jvcir.2019.04.013
14. Zhang T, Zheng W, Cui Z et al (2016) A deep neural network-driven feature learning method for multi-view facial expression recognition. IEEE Trans Multimed 18(12):2528–2536. https://doi.org/10.1109/tmm.2016.2598092
15. Uddin MZ, Khaksar W, Torresen J et al (2017) Facial expression recognition using salient features and convolutional neural network. IEEE Access 5:26146–26161. https://doi.org/10.1109/access.2017.2777003
16. Shao J, Qian Y (2019) Three convolutional neural network models for facial expression recognition in the wild. Neurocomputing 355:82–92. https://doi.org/10.1016/j.neucom.2019.05.005
17. Xie S, Hu H (2019) Facial expression recognition using hierarchical features with deep comprehensive multipatches aggregation convolutional neural networks. IEEE Trans Multimed 21:211–220. https://doi.org/10.1109/tmm.2018.2844085
18. Nguyen H, Kim S, Lee G et al (2019) Facial expression recognition using a temporal ensemble of multi-level convolutional neural network. IEEE Trans Affect Comput. (Early Access Article) https://doi.org/10.1109/taffc.2019.2946540
19. Wang J, Yuan C (2016) Facial expression recognition with multiscale convolution neural network. In: 17th Pacific-rim conference on advances in multimedia information processing, pp 376–385. https://doi.org/10.1007/978-3-319-48890-5_37
20. Hamester D, Barros P, Wermter S (2015) Face expression recognition with a 2-channel convolutional neural network. In: International Joint Conference on Neural Networks (IJCNN). https://doi.org/10.1109/IJCNN.2015.7280539
21. Jung H, Lee S, Yim J et al (2015) Joint fine-tuning in deep neural networks for facial expression recognition. In: IEEE Int Conf Comput Vis (ICCV), pp 2983–2991. https://doi.org/10.1109/iccv.2015.341
22. Happy SL, Routray A (2014) Automatic facial expression recognition using features of salient facial patches. IEEE Trans Affect Comput 6(1):1–12. https://doi.org/10.1109/taffc.2014.2386334
23. Babenko A, Lempitsky V (2015) Aggregating local deep features for image retrieval. In: IEEE Int Conf Comput Vis (ICCV), 1269–1277. https://doi.org/10.1109/iccv.2015.150
24. Lucey P, Cohn JF, Kanade T et al (2010) The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp 94–101. https://doi.org/10.1109/cvprw.2010.5543262
25. Lyons M, Akamatsy S, Kamachi M et al (1998) Coding facial expressions with Gabor wavelets. In: 3rd IEEE International conference on automatic face and gesture recognition, pp 200–205. https://doi.org/10.1109/afgr.1998.670949
26. Taini M, Zhao G, Li SZ, Pietikainen M (2008) Facial expression recognition from near-infrared videos. In: 19th International conference on pattern recognition (ICPR), pp 607–619. https://doi.org/10.1109/ICPR.2008.4761697
27. Dhall A, Murthy OVR, Geoecke R et al (2015) Video and image based emotion recognition challenges in the wild: EmotiW 2015. In: Proceedings of the 2015 ACM on international conference on multimodal interaction, pp 423–426. https://doi.org/10.1145/2818346.2829994
28. Ding H, Zhou SK, Chellappa R (2017) FaceNet2ExpNet: Regularizing a deep face recognition net for expression recognition. In: IEEE 12th International conference on automatic face & gesture recognition, pp 118–126. https://doi.org/10.1109/FG.2017.23
29. Alphonse AS, Dharma D (2017) Enhanced Gabor (E-Gabor), hypersphere-based normalization and pearson general kernel-based discriminant analysis for dimension reduction and classification of facial emotions. Expert Syst Appl 90:127–145. https://doi.org/10.1016/j.eswa.2017.08.013
30. Ming Z, Chazalon J, Luqman MM et al (2018) FaceLiveNet end-to-end networks combining face verification with interactive facial expression-based liveness detection. In: 24th International

conference on pattern recognition (ICPR), pp 3507–3512. https://doi.org/10.1109/ICPR.2018.8545274
31. Viola P, Jones MJ (2004) Robust real-time face detection. Int J Comput Vis 57(2):137–154. https://doi.org/10.1023/B:VISI.0000013087.49260.fb
32. Deng J, Guo J, Ververas E et al (2020) RetinaFace: Single-shot multi-level face localisation in the wild. In: IEEE/CVF Conference on computer vision and pattern recognition (CVPR), pp 5202–5211. https://doi.org/10.1109/CVPR42600.2020.00525
33. Kingma D, Ba J (2015) Adam: A method for stochastic optimization. In: Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015). arXiv:1412.6980
34. Li Y, Zeng J, Shan S et al (2018) Patch-gated CNN for occlusion-aware facial expression recognition. In: 24th International conference on pattern recognition (ICPR), pp 2209–2214. https://doi.org/10.1109/ICPR.2018.8545853
35. Li S, Deng W (2018) Reliable Crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. IEEE Trans Image Proc (TIP) 28:356–370. https://doi.org/10.1109/TIP.2018.2868382
36. Li H, Wang N, Yu Y et al (2021) LBAN-IL: A novel method of high discriminative representation for facial expression recognition. Neurocomputing 432:159–169. https://doi.org/10.1016/j.neucom.2020.12.076
37. Selvaraju RR, Cogswell M, Das A et al (2017) Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: IEEE International conference on computer vision (ICCV), pp 618–626. https://doi.org/10.1109/ICCV.2017.74

Wei Zou received his bachelor degree from Xidian University, China, in 2019. He is currently a postgraduate student in the school of Electronics and Information Technology, Sun Yat-sen University, China. His research interests include deep learning and facial expression recognition.

Dong Zhang received his B.S.E.E. and M. S. degrees from Nanjing University, China, in 1999 and 2003, respectively, and Ph.D. degree from Sun Yat-sen University, China, in 2009. He is currently an associate professor in the school of Electronics and Information Technology, Sun Yat-sen University, China. His research interests include image processing, pattern recognition and information hiding.

Dah-Jye Lee received his B.S. degree from National Taiwan University of Science and Technology in 1984, M.S. and Ph.D. degrees in electrical engineering from Texas Tech University in 1987 and 1990, respectively. He also received his MBA degree from Shenandoah University, Winchester, Virginia in 1999. He worked in the machine vision industry for eleven years prior to joining BYU in 2001. He is currently a Professor in the Department of Electrical and Computer Engineering at Brigham Young University. His research work focuses on object recognition, hardware implementation of real-time vision algorithms and machine vision applications.