Facial Expression Recognition Using Hierarchical Features With Deep Comprehensive Multipatches Aggregation Convolutional Neural Networks

Siyue Xie and Haifeng Hu

***Abstract*—Facial expression recognition (FER) has long been a challenging task in computer vision. In this paper, we propose a novel method, named deep comprehensive multipatches aggregation convolutional neural networks (CNNs), to solve the FER problem. The proposed method is a deep-based framework, which mainly consists of two branches of the CNN. One branch extracts local features from image patches while the other extracts holistic features from the whole expressional image. In the model, local features depict expressional details and holistic features characterize the high-level semantic information of an expression. We aggregate both local and holistic features before making classification. These two types of hierarchical features represent expressions in different scales. Compared with most current methods with single type of feature, the model can represent expressions more comprehensively. Additionally, in the training stage, a novel pooling strategy named expressional transformation-invariant pooling is proposed for handling nuisance variations, such as rotations, noises, etc. Extensive experiments are conducted on the famous the Extended Cohn-Kanade (CK+) dataset and the Japanese Female Facial Expression (JAFFE) database expression datasets, where the recognition results obtained.**

***Index Terms*—Convolutional neural network, expressional transformation-invariant, facial expression recognition, feature aggregation.**

I. Introduction

FACIAL expressions are important carriers for human to convey emotions in communications. Study on nonverbal communication [1] reveals that 55% of a person's emotional or intentional information is conveyed through facial expressions. Recently, researches on emotional analysis have made great achievements. On one hand, development of neuroscience and cognitive science well propel the progress of emotional analysis. On the other hand, technical advance in computer vision and machine learning makes applications related to emotional analysis available to the public. As an important subfield of emotional analysis, researches on Facial Expression Recognition (FER) develop quickly as well. Applications based on FER can be found in many cases, such as human-computer interaction system [2], multimedia [3], surveillance [4] and driver safety [5].

Systematical studies on facial expression analysis can date back to the work of Ekman *et al.* [6]. The main target of facial expression analysis is to establish a system that can automatically classify different expressions. In general, facial expressions can be categorized into six basic expressions [7], which include anger, disgust, fear, happiness, sadness and surprise. Therefore, the primary task of current expressional analysis is to classify these six basic expressions.

Previous methods on FER can be categorized into two groups: detecting facial actions that are related to a specific expression or making classification based on the extracted image features. In Facial Action Coding System (FACS) [8], expressions are encoded by Action Units (AUs), which refer to some tiny but discriminable facial muscle changes. Thus, researchers usually convert FER problem to the task of AU detection. Some other methods represent expressions by some hand-crafted patterns or features, which can be utilized to train an expression classifier. However, some intractable problems are inevitable for these methods. For example, when encoding expressions by AUs, it's hard for researchers to accurately detect every AU in an image as facial muscle moves are sometimes difficult to be tracked. It is also hard to design a type of feature that can adapt to different environments. Variations such as illuminations and image rotations can weaken the representative capacity of hand-crafted features.

In recent years, Convolutional Neural Network (CNN) has achieved great success in the field of computer vision. It can adaptively adjust its convolutional kernels of each layer to obtain some desired features, which makes it adapt to various classification problems without much prior knowledge. In general, the output of CNN, i.e., the feature maps of the last layer, will be treated as the high-level semantic concept to represent the input. In FER problem, previous studies [9], [10] have revealed that expressional changes usually occur on some salient facial regions such as neighborhood of mouth, eyes and

Manuscript received June 3, 2017; revised February 1, 2018 and March 23, 2018; accepted May 5, 2018. Date of publication June 4, 2018; date of current version December 20, 2018. This work was supported in part by the National Natural Science Foundation of China under Grants 61673402, 61273270, and 60802069; in part by the Natural Science Foundation of Guangdong under Grants 2017A030311029, 2016B010109002, 2015B090912001, 2016B010123005, and 2017B090909005; in part by the Science and Technology Program of Guangzhou under Grants 201704020180 and 201604020024; and in part by the Fundamental Research Funds for the Central Universities of China. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Ivan V. Bajic. *(Corresponding author: Haifeng Hu.)*

The authors are with the School of Electronic and Information Engineering, Sun Yat-sen University, Guangzhou 510275, China (e-mail: xiesy8@mail2.sysu.edu.cn; huhaif@mail.sysu.edu.cn).

Digital Object Identifier 10.1109/TMM.2018.2844085

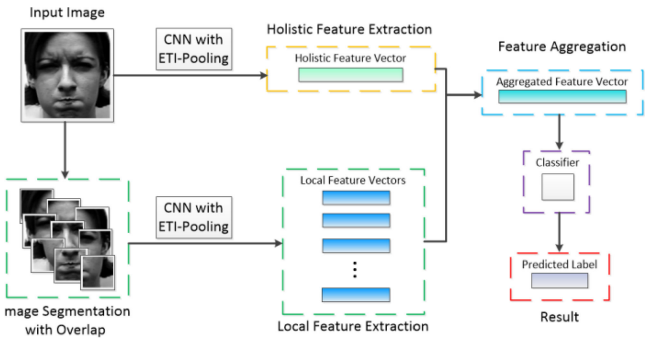


Fig. 1 Flowchart of the proposed method.

nose. This implies that details of local facial regions can be discriminative for expressional recognition. However, most current CNN-based methods on FER only extract features from the whole expressional image. These methods emphasize the integrality of a facial expression but ignore the information of local details. In the other words, typical CNN may not take full advantage of recognition-effective information encoded in expressional images. Therefore, we attempt to make some modification in commonly used CNN architecture for improving performance on FER problem.

In this paper, we present a novel framework, named Deep Comprehensive Multi-Patches Aggregation Convolutional Neural Networks (DCMA-CNNs), for solving the FER problem. Fig. 1 illustrates the flowchart of our method. The framework consists of two individual CNN branches: one extracts holistic features from the whole image and the other extracts local features from some overlapped image patches. Holistic features aim to represent the integrity of an expression while local features focus on describing details on local regions, which can indirectly indicate some active expressional regions on the face. These two separated branches represent an expressional image from two different scales, each of which is complementary to the other. Compared with current works, which mostly characterize expressions using single type of features, DCMA-CNNs can represent an expressional image more comprehensively by aggregating two types of hierarchical features. In addition, we improve the TI-pooling [11] and propose the Expressional Transformation-invariant pooling (ETI-pooling) strategy for handling nuisance variations such as illuminations, image rotations and noises. ETI-pooling enhances the discriminative ability of our model by fusing a certain number of features that from the same class. This pooling strategy is applied to modify the structure of typical CNN and well improve the recognition performance. Extensive experiments are conducted on the CK+ and JAFFE dataset and classification results demonstrate the effectives of the proposed method.

Contributions of our model can be summarized as follow:

1) A novel two-branch framework is proposed to solve the challenging FER problem. One branch is implemented to extract local features from image patches, which highlights the detailed information of facial expressions. Salient patches and active regions related to FER problem can be determined and visualized based on the extracted local features.
2) We incorporate both the holistic and local features into our model. These two types of hierarchical features represent images in different scales. By aggregating these features, we can obtain a more discriminative representation of an image, which significantly improves the classification performance.
3) We modify typical CNN structure with the proposed ETI-pooling, which can distinguish expression-sensitive elements from extracted features. Owing to the modification, our model can be more robust to some variations such as illuminations, image rotations, noises, etc.

The remainder of this paper is organized as follow. We make a brief review of some existing works on FER in Section II. The details of DCMA-CNNs will be specialized in Section III. In Section IV, experiments and result analysis are presented. We conclude our method in the Section V.

II. Related Works

Expression recognition has long been a challenging problem in emotional analysis. Conventional approaches can be generally categorized into two groups: AU-based and feature-based approaches. Recently, methods based on deep-learning algorithms make great achievements in expression recognition, which is regarded as an effective alternative to solve FER problem.

AU-based methods attempt to detect expression-related AUs on a facial image, which are inspired by the studies on FACS [12]. Researchers can recognize an expression according to the combination of detected AUs [13]–[17]. Some other methods even focus on 3D AU detection [18], [19]. As for feature-based method, hand-crafted patterns or features are usually used to represent an expressional image. Geometric relationships among facial organs or landmark points are typical features to represent expressions [20]–[23]. Some image operators, such as Local Binary Patterns (LBP) [24], Local Description Patterns (LDP) [25], Gabor-features [26] Local Phase Quantization (LPQ) [27], [28] or Scale Invariant Feature Transform(SIFT) [29], are used in expression analysis as they can extract significant information from images. Some methods divide images into patches with different scales. Features extracted from patches can highlight some detailed information of local facial regions [30], [31], [9], which significantly improve the recognition performance. To enhance the comprehensive representation ability, methods like [10] even fuse different types of features before making classification.

In recent years, deep-based algorithms have been applied to expression recognition. Zhao *et al.* [32] propose the Deep Region and Multi-label Learning (DRML) algorithm to conduct AUs detection, which addresses the problem of Region Learning (RL) and Multi-label Learning (ML). Li *et al.* [33] extract Gabor-wavelet features from images to train a deep network for FER task. Mollahosseini *et al.* [34] construct a CNN with inception layers. Their experiments are conducted on seven standard face datasets and obtained comparable results. Liu *et al.* [35] model a system named Boosted Deep Belief Network (BDBN) to classify different expressions. They divide expressional images into patches. Some patches with high discriminative power will be selected and combined to train a strong classifier.

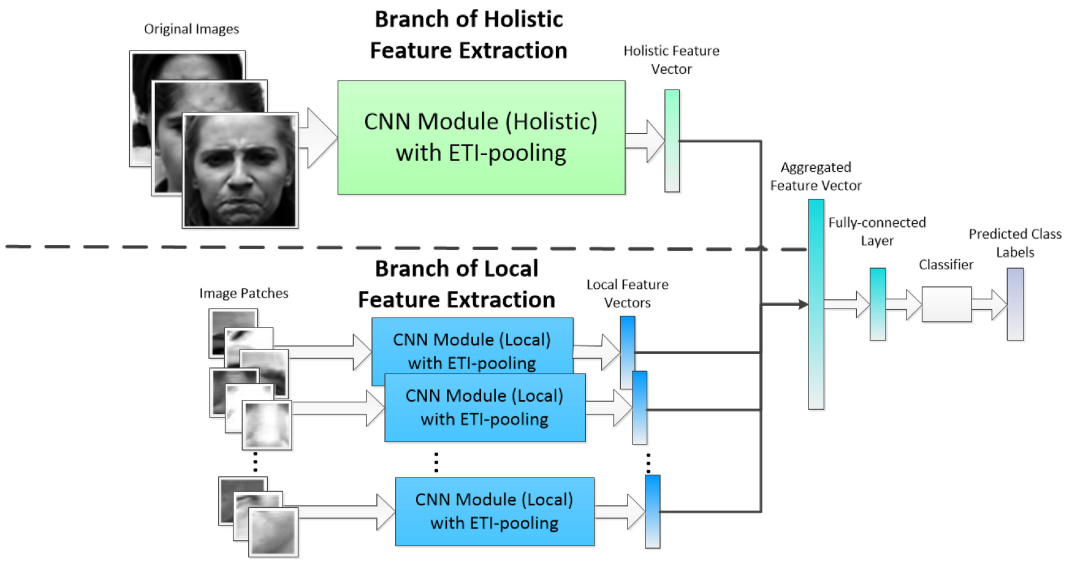


Fig. 2 Framework of DCMA-CNNs.

Some works aggregate or fuse different types of features through deep network to produce a comprehensive representation, which usually results in a better recognition result than using single type of features. For example, Majumder *et al.* [10] extract LBP features as well as facial geometrical features from expressional images. These two types of features are finally fused by a 2-layer autoencoder. Hamester *et al.* [36] construct 2-channel architecture for feature extraction, which utilizes CNNs and an autoencoder to extract features. Jung *et al.* [37] train a CNN and a deep neural network independently and combine them through fully connected layers with joint fine-tuning. However, it should be noted that most existing deep-based methods merely focus on extracting high-level semantic concept of expressions but ignore fine-grained information in local facial regions. Different from some existing works, in this paper, we present an individual CNN branch to extract local features, which highlights the importance of local detailed information in expression analysis.

III. Proposed Method

The proposed model is based on a two-branch CNN framework, as shown in Fig. 2. In order to extract variation-robust features, we propose the ETI-pooling strategy to modify the typical structure of CNN, which forms the CNN module. A CNN module is a basic unit for feature extraction in the proposed DCMA-CNNs model, and it helps extract local and holistic features from an input image.

In the following subsections, we briefly review the structure of typical CNN (Section III.A) and then describe the special case of obtaining the desired CNN module by adopting ETI-pooling (Section III.B). The obtained CNN module will be used to construct the DCMA-CNNs model (Section III.C). Additionally, a salient patch learning algorithm is presented based on the proposed DCMA-CNNs model (Section III.D). This algorithm can indicate some active facial regions that are relevant to changes in expression.

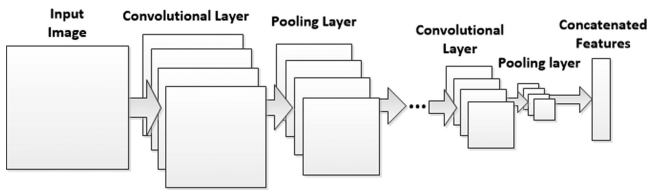


Fig. 3 A typical structure of CNN.

A. Convolutional Neural Network

CNN has been successfully applied in many tasks related to computer vision in the recent years. Typically, a CNN comprises stacks of multiple convolutional layers and pooling layers. In general, the features generated by the deeper layers can represent the content of a larger region in the input image. Feature maps yielded by the last layer can be treated as the representation of the input. The typical structure of a CNN is shown in Fig. 3.

In our model, the CNN module has a four-layer network structure. Convolutional layers and pooling layers are stacked up alternately to forms the backbone of the CNN module. The output feature maps in the final layer serve as the high-level semantic concept of the input, and they are used for the processing as described in the following subsections.

B. CNN With Expressional Transformation-Invariant Pooling

In the stage of feature extraction, variations such as illumination or image rotation can degrade the representative capacity of the extracted features. One approach commonly used to address this problem is dataset augmentation. However, this approach has some side-effects. Models trained with data augmentation still need to learn feature representations separately for different variations. Extra transformation of data can mislead models to learn some useless information from noises samples or wrong labels [11].

In order to reduce the negative impact of image variations and to overcome the disadvantages of data augmentation, we

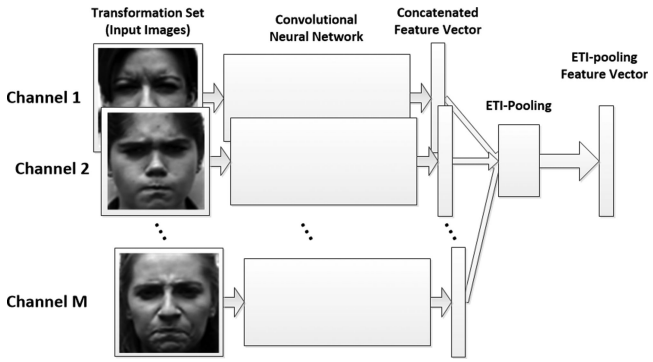


Fig. 4 Modified CNN structure with ETI-pooling (a CNN module).

introduce a novel feature-fusing strategy, namely, Expressional Transformation-invariant pooling (ETI-pooling), to modify the typical CNN structure. It follows the concept of TI-pooling [11], which was proposed to handle one specific variation in images. TI-pooling accumulates all features of input samples and only responses to the maximal feature. Such processing allows for more efficient data usage and can help discover "canonical" instances, leading to the generation of variation-independent and transformation-invariant features.

Unlike TI-pooling, our ETI-pooling scheme was extended to handle different variations simultaneously. In our method, different variations can be regarded as different transformations for images in the same class. For example, for two images with expression of anger but different illuminations, one image can be regarded as an illumination transformation version of the other. We can accumulate the responses of all samples from the same class instead of treating them individually. Such processing not only allows TI-pooling to learn from "canonical" instances, but also yields features that are robust to different variations.

The main structure of ETI-pooling is illustrated in Fig. 4. To extract invariant features from expressional images, we train the network using samples from the transformation sets, which consist of a certain number of expressional images with the same class label. In the training stage, transformation sets are treated as the input of the networks. Each image of a transformation set will be fed into a channel, i.e., an expressional image corresponds to a channel of the CNN module. In each channel, the input image will pass through a CNN, whose interior structure is similar to that shown in Fig. 3. The output of each channel comprises vectorized features that are denoted via a concatenated feature vector (refer to Fig. 3). In the training stage, weight-sharing is implemented among these parallel channels, indicating that the model only requires the same amount of memory as one CNN. The output features of every individual channel will finally be fused to generate a feature vector.

We denote the concatenated feature vector of the k-th ($k = 1, 2, \ldots, M$) channel as $\boldsymbol{z}^k = (z_1^k, z_2^k, \ldots, z_t^k, \ldots, z_N^k)$, where N is the dimension of the vector and $z_t^k (t = 1, 2, \ldots, N)$ is the t-th element of the vector. The vector obtained after ETI-pooling can be denoted as $\boldsymbol{z} = (z_1, z_2, \ldots, z_t, \ldots, z_N)$. The non-linear operator conducts an element-wise fusion operation across all M channels, formulated as follows:

$$z_t = \max\left(z_t^1, z_t^2, \ldots, z_t^M\right) \tag{1}$$

This operation only responds to the maximal element in the same position among different feature vectors, which lowers the probability of propagating an odd variation-related feature to the following parts. In the other words, ETI-pooling attempt to improve the classification performance by highlighting the most discriminative feature elements that are beneficial to most instances. Such non-linear operations increase the competition among different channels, inducing networks to adaptively learn the expression-discriminative features and to suppress the back propagation of information related to nuisance variations. Unlike the max-pooling strategy, ETI- pooling will not reduce the dimension of a representation. This makes it efficiently utilize all available and effective information concealed in multiple expressional images. By training the network using transformation sets, the network can learn invariant features across variations. This is beneficial to classification as described below.

C. DCMA-CNNs for Classification

Many existing deep-based models for facial expressional analysis are built on single-branch CNN [33], [34], [38], which usually focuses on extracting features from a whole image but ignores detailed descriptions. However, some studies [9], [22], [30] validated the effectiveness of local expressional features in solving the FER problem. This means that both holistic and local features are important for expressional analysis. Some recent methods fuse features from different representative spaces [10], [36], [37] or aggregate different features [39], [40] that achieved satisfactory performance in different classification tasks. It is reasonable that fused or aggregated features contain more classification-effective information than a single type of features. Motivated by these works, we attempted to incorporate both holistic and local information and aggregate them for improving the representation ability of our model.

As shown in Fig. 2, the DCMA-CNNs consist of two branches. One extracts holistic features from input images, while the other focuses on local feature extraction. The output features obtained from the two branches are aggregated and used in expression recognition task.

In the branch of holistic feature extraction, images from a transformation set are treated as the input of the CNN module. The output of the branch is the holistic feature vector, which represents a set of images with different transformations. In the other branch, original images are equally divided into image patches. Individual CNN modules are implemented to extract the local features with respect to each patch. All the steps of local feature extraction are the same as those of the holistic feature extraction except for some detailed configuration of CNN modules. Although both branches are based on CNN, the difference in input data results in complementary output features with different semantic concepts. The local branch focuses on extracting patch-specific semantic concepts, while the holistic branch concentrates on extracting an abstract representation from the whole image. Aggregation of these two branches not only increases the number of extracted features, but also improves the representative ability of our model.

In this paper, we denote the holistic feature vector as $\boldsymbol{v}_h^0$, and the m-th local feature vector as $\boldsymbol{v}_l^m$, where $m = 1, 2, \ldots, L$,

L is the number of local feature vectors, which equals to the number of image patches. The aggregated feature vector $\boldsymbol{v}_a$ is obtained by concatenating holistic feature vector and all local feature vectors, which can be formulated as follows:

$$\boldsymbol{v}_a = \left(\boldsymbol{v}_h^0; \boldsymbol{v}_l^1; \boldsymbol{v}_l^2; \ldots; \boldsymbol{v}_l^m; \ldots; \boldsymbol{v}_l^L\right) \tag{2}$$

The aggregated feature vector will be fed into a fully-connected layer for feature fusion. A classifier is followed by the fully-connected layer, which maps the output of the previous layer to an expression class. In our method, we choose softmax as the classifier, which can be formulated as shown below:

$$f\left(\boldsymbol{z_j}\right) = \frac{e^{\boldsymbol{z_j}}}{\sum_i e^{\boldsymbol{z_i}}} \tag{3}$$

Each element of equation (3) corresponds to a unique expression class. The class with the maximal element value is regarded as the predicted class. Thus, the loss function of our model can be formulated as follows:

$$L = \frac{1}{2}\|\boldsymbol{y^p} - \boldsymbol{y^g}\|^2 = \frac{1}{2}\sum_{c=1}^{C}\left(y_c^p - y_c^g\right)^2 \tag{4}$$

where $\boldsymbol{y^p}$ is the predicted label; $\boldsymbol{y^g}$ is the ground truth label; y_c^p and y_c^g are the c-th element of $\boldsymbol{y^p}$ and $\boldsymbol{y^g}$ respectively. C is the total class number of expressions. All related parameters of the method can be updated through back-propagation using the Stochastic Gradient Descent (SGD).

D. Learning Salient Patches for Expression Analysis

Intuitively, expressions can be regarded as regional changes in the appearance where the changes are triggered by some facial muscles. This means that analysis of expressions can be associated with limited facial regions. Some previous works on FER are based on the study of local regions [9], [10], [30]. In this subsection, we define the salient patches as the image patches that make large contribution to the expression recognition task. As local regions can be represented by the extracted local features, we can detect these salient patches by analyzing the features extracted from them.

In classification tasks, features with large norms are much more discriminative than random features [39]. Therefore, we can try to detect salient patches according to the L2-norm value of the extracted local feature vectors. We assume that the discriminative power of a local feature vector is proportional to the magnitude of its L2-norm.

Based on this assumption, we can bypass some patches with a small norm but preserve patches with a large norm. The procedure of learning salient patches is implemented in two steps in the training stage: For the first step, all feature vectors are involved in the classification task. As the training loss converges, the second step is initiated to learn the salient patches and to continue fine-tuning the model. The general procedure of selecting salient patches is summarized in Algorithm 1. We record the norm of each local feature vector in each iteration and compute the Averaged Cumulative Summation (ACS) of the norm of each vector every T iterative epochs. According to our assumption, vectors with larger ACS play a more important role in recognition. The vector with the smallest ACS contributes the least; in other words, we can bypass this patch and exclude it in the next iteration when continuing fine-tuning. By successively removing irrelevant patches from the input, we can optimize the model and increase the representation efficiency.

Algorithm 1: Leaning Salient Patches for Expression Analysis

Input: Localized feature vectors set $\Phi = \{\mathrm{v}_1, \mathrm{v}_2, \ldots, \mathrm{v}_\mathrm{m}, \ldots, \mathrm{v}_\mathrm{N}\}$ corresponds to N patches, Salient patches number P, bypass period T

1: Initialize averaged cumulative norm summation set $\Omega = \{\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_m, \ldots, \boldsymbol{s}_N\}$ for N patches, $\boldsymbol{s}_m = 0, m = 1, 2, \ldots, N$;

2: Begin training until the training loss converges;

3: **While** *number of elements in* Φ *has not met* P, **do**

  for $t = 1 : T$ **do**

    Fine-tune the model with localized feature vectors in Φ;

    for each *element in* Ω **do**

      $\boldsymbol{s}_m \leftarrow \boldsymbol{s}_m + \frac{1}{T}\|\boldsymbol{v}_m\|^2$;

    end for each

  end for

  $k = \arg\min_m \boldsymbol{s}_m$;

  $\Phi \leftarrow \Phi - \boldsymbol{v}_k$;

  $\Omega \leftarrow \Omega - \boldsymbol{s}_k$;

  end while

4: **Output**: Localized feature vectors set Φ

IV. Experimental Evaluation

In this section, a number of experiments are carried out on two publicly available facial expression datasets: the Extended Cohn-Kanade (CK+) dataset [41] and the Japanese Female Facial Expression (JAFFE) database [42]. Experiments are in support of the following objectives:

1) Evaluate the performance of DCMA-CNNs on FER task by the recognition accuracy and compare it with some competing conventional approaches as well as deep-based methods.
2) Investigate the role of aggregated features and ETI- pooling in solving FER problem.
3) Investigate the effect of salient patch learning algorithm and the impact on recognition accuracy after bypassing some irrelevant patches.

A. Experimental Data

Experiments are conducted on CK+ and JAFFE datasets. For each dataset, six classes of basic expressions, i.e., anger, disgust, fear, happiness, sadness and surprise, are chosen as the target of our classification task.

The CK+ dataset contains 327 expression-labeled image sequences from 123 subjects. All sequences are from the neutral face to the peak expression. Some samples of the CK+ dataset are shown in Fig. 5(a). 309 sequences with one of the six prototypical expressions are selected to conduct experiments. For each chosen sequence, the last 3 frames with peak expression are

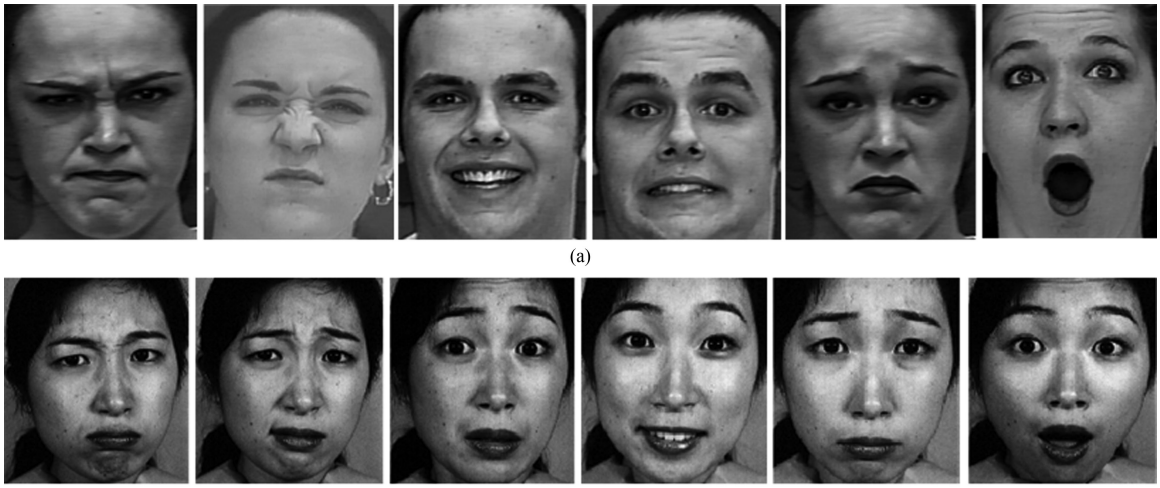


Fig. 5 Example of the six prototypic expressions. (a) The CK+ dataset. (b) The JAFFE database. (left to right: anger, disgust, fear, happiness, sadness, surprise).

TABLE I
NETWORK CONFIGURATION (CONV: CONVOLUTIONAL; MP: MAX-POOLING; FC: FULLY-CONNECTED)

	Holistic Branch	Localized Branch
	Type / kernel size	Type / kernel size
Layer1	Input / -	Input / -
Layer2	Conv / $5 \times 5 \times 6$	Conv / $3 \times 3 \times 6$
Layer3	MP / 2×2	MP / 2×2
Layer4	Conv / $5 \times 5 \times 18$	Conv / $3 \times 3 \times 18$
Layer5	MP / 2×2	MP / 2×2
Layer6	FC / 8424×6	

collected to form the training and testing sets, i.e., 927 images are involved in our experiments.

The JAFFE database contains 213 images of 7 facial expressions posed by 10 Japanese females. 183 images with six basic expressions are chosen for the experiment. Fig. 5(b) shows some samples of the JAFFE database. As for JAFFE dataset, data augmentation is applied for improving recognition performance. Images are horizontally flipped, which obtains the corresponding mirror image. Each image is rotated by the angles of 5° clockwise and counterclockwise. Furthermore, Gaussian noise with a zero mean and 0.01 variance is added to the original image. Thus, the sample set of experiments on JAFFE database is largely extended, which contains 915 facial expression images.

To simplify the recognition task, human faces in all selected images are detected by the Viola-Jones faces detector [43]. Facial regions are cropped from images and resized to the scale of 60×60. These resized facial expression images are then equally divided into partially overlapped patches, which forms the sample set of the branch of local feature extraction.

B. Experimental Settings

In our experiment, two individual branches are set with different CNN structural configuration. Table I shows the detailed configuration of the CNN module in each branch. The stride of the convolutional layer is set as 1. The convolutional operation is conducted without padding. When applying the ETI-pooling, we set 3 parallel channels in the model.

In the training stage, the learning rate α decays in an exponential form, which can be formulated as:

$$\alpha = \alpha_0 \times 0.95^{\lfloor \frac{i}{10} \rfloor} \tag{5}$$

where $\alpha_0 = 0.1$ is the initial value, i indicates the i-th iterative epoch of the training stage and symbol $\lfloor x \rfloor$ refers to the largest integer that is smaller than x. In our experiment, images are divided into 3×3 patches with overlap. The overlap ratio between two adjacent patches is 0.5.

To evaluate the performance of the proposed method, 10-fold cross-validation is applied to all the experiments (i.e., images are randomly divided into ten equal-sized subsets, nine for training and one for testing). To fairly compare with other methods, in CK+, all ten subsets are subject-independent. As for the JAFFE database, an original image and its augmented images can only be allocated either to training or testing sets. The final results are reported by averaging the recognition accuracy of all ten folds experiments.

C. Expression Recognition Results

1) Experiments on the CK+ Dataset: In order to show the detailed performance of our model, we list the confusion matrix of each expression class in Table II. Moreover, we compare our model with many existing works and Table III shows the comparison results. In this paper, we compare the recognition accuracy of our model with several competitive approaches, including conventional methods (LAP [44], MPSD [45] and MCSPL [30]) and deep-based methods (DTAN [37], DTGN [37] and 3DCNN-DAP [46]). In addition, we reproduce the model of the work of Khorrami *et al.* [47], which is a recently-proposed deep-based model and is denoted as Khor-net in the following.

From Tables II and III, we can observe the following.

1) The proposed method achieves an averaged recognition accuracy of 93.46% in the CK+ dataset. DCMA-CNNs performs the best when recognizing the expression of sur-

TABLE II
CONFUSION MATRIX OF DCMA-CNNS ON CK+ (%) (AN: ANGER, DI: DISGUST, FE: FEAR, HA: HAPPINESS, SA: SADNESS, SU: SURPRISE)

	An	Di	Fe	Ha	Sa	Su
An	90.74	0	0	0	9.26	0
Di	1.85	95.06	1.85	0	1.23	0
Fe	11.11	0	74.07	3.70	0	11.11
Ha	0	2.12	3.17	94.71	0	0
Sa	12.35	1.23	1.23	0	81.48	3.70
Su	1.39	0	0.46	0.46	0	97.69

TABLE III
PERFORMANCE COMPARISON WITH EXISTING WORKS ON THE CK+ DATASET (%)

Method	Accuracy
LAP [44]	88.26
MPSD [45]	88.52
Khor-Net [47]	91.25
MCSPL [30]	91.53
DTAN [37]	91.44
DTGN [37]	92.35
3DCNN-DAP [46]	92.40
DCMA-CNNs (proposed)	93.46

prise. Some samples of fear and sadness are misclassified as anger. This is because expressions of fear and anger have some similar actions in local facial regions, such as fall of eyebrow and rise of the upper eyelid. In FACS, these two expressions have shared AUs. The same phenomenon appears when dealing with sadness and anger. This may be due to the reason that only 75 samples of fear and 84 samples of sadness are used for training, which is far less than that of other four expressions.

2) The performance of DCMA-CNNs is far better than these conventional methods. LAP and MCSPL extract LBP features to represent expressions while MPSD utilizes SIFT features. All these features are artificially predefined and incapable to be adaptive to some complex environments. Compared with these methods, our deep-based model learns features adaptively, which are more robust than hand-crafted features.
3) DCMA-CNNs outperforms the competitive deep models. The deep-learning based methods such as DTAN and DTGN utilize one-branch architecture to extract desired features while our method extracts features with an additional branch, which leads to a more comprehensive representation of expressions. The result indicates that the branch of local feature extraction really benefits expression classification.
4) DCMA-CNNs outperforms Khor-net. This is because Khor-net is based on the structure of typical CNN while our model adopt ETI-pooling and feature aggregation schemes. Thus our DCMA-CNNs model can capture more detailed information and is robust to different variations. Note that the recognition accuracy of Khor-net in Table III is different from its reported result in [47]. Although we use the same dataset (CK+), validation settings (10-fold cross-validation) and network architecture as [47], images used in each fold of experiments can be different as training and testing samples are randomly selected from dataset. Such difference results in the fluctuation of the final recognition performance.

TABLE IV
CONFUSION MATRIX OF DCMA-CNNS ON JAFFE (%)

	An	Di	Fe	Ha	Sa	Su
An	97.04	2.22	0	0	0	0.74
Di	3.97	93.65	0.79	0	0.79	0.79
Fe	2.78	1.39	88.89	0.69	1.39	4.86
Ha	0.69	0	0.69	95.14	0.69	2.78
Sa	2.22	3.70	2.22	0	90.37	1.48
Su	0	0.74	1.48	2.22	0	95.56

TABLE V
PERFORMANCE COMPARISON WITH EXISTING METHODS ON JAFFE (%)

Method	Accuracy
McFIS [48]	87.60
LPTP [49]	90.20
SRC+LBP [50]	90.30
SRC+Gabor [50]	91.21
SRC+LPQ [50]	91.67
SFPL [9]	91.80
LSDP [51]	92.30
DCMA-CNNs (proposed)	94.75

2) Experiments on the JAFFE Database: To validate the performance of DCMA-CNNs on the JAFFE database, we show the detailed classification results of each expression class in Table IV. We also compare our model with many other methods by recognition accuracy. The comparison results are listed in Table V.

From Tables IV and V, some conclusions can be summarized as the following.

1) The proposed method achieves an averaged recognition accuracy of 94.75% in the JAFFE database. The best averaged recognition accuracy occurs when recognizing the expression of anger, which reaches the accuracy of 97.04%. The averaged recognition accuracy on fear is relatively lower than all others but still reach 88.89%. In the experiment, nearly 5% samples of fear expressions are misclassified as surprise. A reasonable explanation is that the expression of fear and surprise act similarly in several facial regions such as the inner brows and the upper eyelids. We can also find that these two expressions are coded by some identical AUs in FACS.
2) The proposed DCMA-CNNs model outperforms all other competitive methods. In Table V, some methods using hand-crafted features are involved in comparison but the recognition accuracy of them is far lower than the proposed model, which validates the effectiveness of deep-based features.
3) The proposed method is robust to small variations. Images with variations such as rotations and noises are involved in the experiment on JAFFE but DCMA-CNNs can still

TABLE VI
GENERALIZATION PERFORMANCE OF DM-CNNS ACROSS DATABASES (TRAINED ON JAFFE, VALIDATED ON CK+) (%)

	Accuracy
da Silva *et al.* [52]	48.20
DCMA-CNNs	46.28

TABLE VII
RECOGNITION ACCURACY ON CK+ WITH DIFFERENT TYPES OF FEATURES (%)

	Accuracy
DCMA-CNNs	93.46
HFCNN	90.67
LFCNN	88.67

correctly classify most expressions, which demonstrates the robustness of the proposed method.

3) Experiments Across Databases: To evaluate the generalization capability of our model, we conduct an experiment on DCMA-CNNs across different databases. To be specific, we train DCMA-CNNs with the data from the JAFFE databases and validate it on the CK+ dataset.

Note that there are few works had conducted cross-databases testing on facial expression recognition. It is hard to make a full and fair comparison with other models. Here we compared our model with the work of da Silva *et al.* [52]. To make a fair comparison, our experimental settings and preprocessing follow the work of da Silva *et al.* [52]. The results are shown in Table VI. From the table, we can observe that DCMA-CNNs is still comparable to the work of [52], which demonstrate the effectiveness of our model.

D. Comparison Studies on DCMA-CNNs Properties

In our method, we aggregate two types of features and introduce the ETI-pooling into the model. To assess these two properties, we conduct some experiments on the CK+ dataset to evaluate their effect on recognition.

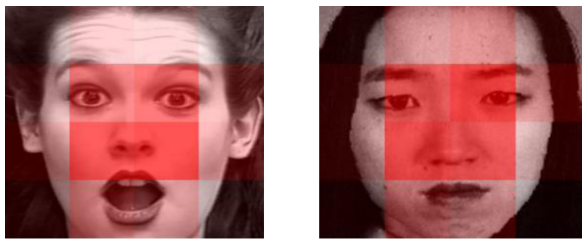
1) The Effects of Feature Aggregation: We construct another two models to take on the evaluation task. The model only utilizes holistic features to make classification is denoted as HFCNN. The model that recognizes expressions only with local features is denoted as LFCNN. The recognition performances of these two models are listed in Table VII.

From Table VII, we can observe that the recognition accuracy of DCMA-CNNs is much higher than HFCNN and LFCNN, which means that feature aggregation really improve expression recognition. This is reasonable as holistic features or local features only focus on representing expressional information with a specific scale. The improvement on recognition accuracy by aggregation indicates that these two kinds of features are complementary to each other.

2) The Effects of ETI-Pooling: In this experiment, the model without ETI-pooling is denoted as DCMA-CNNs-WTE. In other words, in this model, the CNN module is replaced by a typical CNN structure as shown in Fig. 3. We compare the performance of DCMA-CNNs-WTE with the proposed model. The recognition result is shown in Table VIII.

TABLE VIII
EFFECT OF ETI-POOLING ON CK+ DATASET (%)

	Accuracy
DMA-CNN	93.46
DMA-CNN-WTE	92.67



(a) (b)

Fig. 6 Visualization of significant patches and the corresponding active regions following the learning strategy of salient patches. The intensity of red color is proportional to the total times a region is retained in experiments. (a) Example on the CK+ dataset. (b) Example on the JAFFE database.

From Table VIII, we can see that the proposed model performs better than DCMA-CNNs-WTE. This can be owed to the parallel structure and the non-linear operator of ETI-pooling. These two operations help distinguish discriminative elements in the feature vector and suppress nuisance variations collaboratively. The improvement in recognition accuracy demonstrates the effectiveness of ETI-pooling.

E. Experiments on Learning Active Regions

Following the assumptions and the learning algorithm we suggested in the Section III.D, we attempt to make a selection of expressional patches in our experiments. The period of selection is 30 epochs in the experiments on both the CK+ dataset and JAFFE database, i.e., we bypass an image patch every 30 iterative epochs during the procedure of fine-tuning. Experiments follow the principle of 10-fold cross-validation and we finally retain 5 patches in the end of each fold of experiment. We count the times that a patch is retained after all ten folds of experiments. The more times a patch is retained, the more importantly it acts in the expressional analysis. Note that an expressional image is divided by the scale of 3×3 with 50% overlap area between two adjacent patches, an image can be seen as being equivalently divided into 4×4 disjoint regions in the visualization task. Each patch is composed of four regions. Regions in the retained patch can be regarded as active regions whereas regions in bypassed patches are regarded as less-active regions.

We visualize the salient patches and their corresponding active regions in Fig. 6. The intensity of red color of a region is proportional to the times it is retained in the 10-fold cross-validation experiments. We can observe that in both the CK+ dataset and the JAFFE database, the "most active" regions are located around some primary facial organs such as eyes, nose and mouse. Some marginal regions are with low intensity, which means that these "less-active" regions and their corresponding patches contribute less to expression recognition. Distribution of active regions roughly accord with the intuition of places where actions of an expression should take place in face.

TABLE IX
RECOGNITION ACCURACY OF DMA-CNN WITH AND WITHOUT PATCHES EXCLUSION IN CK+ AND JAFFE (%)

	CK+	JAFFE
DCMA-CNNs	93.46	94.75
DCMA-CNNs-5p	92.10	94.75

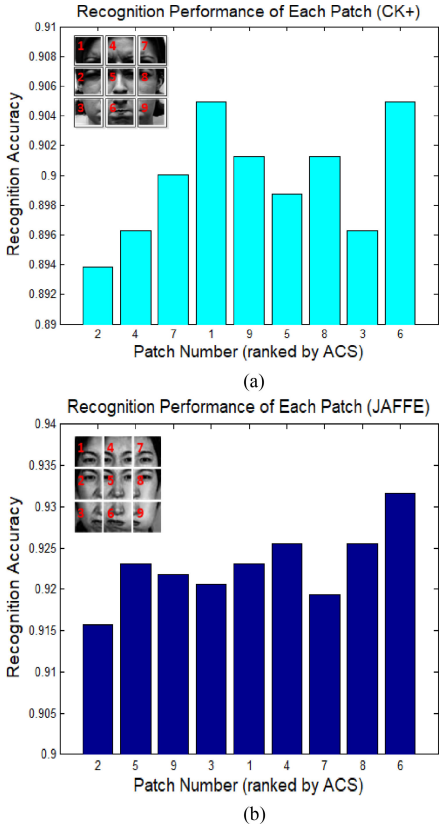


Fig. 7 Recognition Performance with respect to different patches. (a) Performance on the CK+ dataset. (b) Performance on the JAFFE database.

The recognition performance after bypassing less-active patches is listed in Table IX. We denote DCMA-CNNs-5p as the model trained with five retained patches. We can observe that the averaged recognition accuracy of DCMA-CNNs-5p in the CK+ dataset is slightly lower than the accuracy of DCMA-CNNs. This is reasonable as DCMA-CNNs can make use of more local information for making classification. However, these two models achieve the same averaged recognition accuracy in the JAFFE database, which implies that the retained five patches have contained most discriminative information of expressions. To sum up, the proposed model and the salient patch learning algorithm help distinguish sensitive patches that are related to the FER problem.

We conduct another experiment to verify the assumption mentioned in Section III.D, i.e., patches with larger ACS plays a more important role in FER. To be specific, we prune the local branch of DCMA-CNNs. Each pruned DCMA-CNNs can extract features from only one specific patch. Thus we can totally construct nine different pruned models corresponding to each different patch. The recognition performance of each network is shown in Fig. 7, where we sort all patches by the magnitude of ACS (in ascending order). From the figure, we can observe that models with the patch that have larger ACS tend to have a better recognition performance. In both the CK+ and JAFFE dataset, the model with the largest patch ACS achieves the highest recognition accuracy. This verifies our assumption that patches with larger ACS can be more important in FER task.

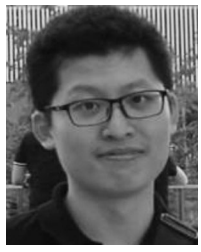
V. CONCLUSIONS

In this paper, a novel method, named DCMA-CNNs, is proposed for facial expression recognition. The architecture of our proposed model consists of two individual CNN branches. One of the branches conducts holistic features extraction on the whole expressional images while the other extracts local features from segmented expressional image patches. These two types of hierarchical features depict an expressional image in two different scales and being complementary to each other. Expressional images represented by these two types of features can be more discriminative than many existing works. We aggregate the holistic and local features to yield fused features and classification is conducted based on the fused features. Furthermore, we modify the typical CNN structure with the proposed ETI-pooling strategy, which reduces the impact of nuisance variations in classification tasks. We additionally proposed a method to learn salient expressional image patches based on the L2-norm of local feature vectors and visualize the active regions relevant to expression changes. The selected salient patches are regarded as important parts for facial expression recognition. Extensive experiments on two publicly available expression datasets (the CK+ dataset and the JAFFE database) demonstrate the effectiveness of our proposed method on FER task. The classification result of our DCMA-CNNs model on these two datasets outperforms many other competitive works, which include both conventional and deep based methods.

REFERENCES

[1] G. Mehrabian, *Nonverbal Communication*. New Brunswick, NJ, USA: Aldine, 2007.
[2] A. Ryan *et al.*, "Automated facial expression recognition system," in *Proc. 43rd Annu. Int. Carnahan Conf. Security Technol.*, Zurich, Switzerland, 2009, pp. 172–177.
[3] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image Vis. Comput.*, vol. 31, no. 1, pp. 1743–1759, 2009.
[4] Q. Wang, K. Jia, and P. Liu, "Design and implementation of remote facial expression recognition surveillance system based on PCA and KNN algorithms," in *Proc. Int. Conf. IEEE Intell. Inf. Hiding Multimedia Signal Process.*, Adelaide, Australia, 2015, pp. 314–317.
[5] E. Vural *et al.*, "Automated drowsiness detection for improved driver-safety comprehensive databases for facial expression analysis," in *Proc. Int. Conf. Autom. Technol.*, vol. 1. Istanbul, Turkey, 2008, pp. 96–105.
[6] P. Ekman and W. Friesen, *The Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Santa Clara, CA, USA: Consulting Psychologists Press, 1978.
[7] C. E. Izard, *The Face of Emotion*, vol. 1. New York, NY, USA: Appleton-Century-Crofts, 1971.
[8] P. Ekman, W. V. Friesen, and J. C. Hager, "FACS manual," Salt Lake City, UT, USA: A Human Face, May 2002.
[9] S. Happy and A. Routray, "Automatic facial expression recognition using features of salient facial patches," *IEEE Trans. Affect. Comput.*, vol. 6, no. 1, pp. 1–12, Jan.–Mar. 2015.
[10] A. Majumder, L. Behera, and V. K. Subramanian, "Automatic facial expression recognition system using deep network-based data fusion," *IEEE Trans. Cybern.*, vol. 99, pp. 1–12, Jan, 2016.

[11] D. Laptev, N. Savinov, J. M. Buhmann, and M. Pollefeys, "Ti-pooling: Transformation-invariant pooling for feature learning in convolutional neural networks," in *Proc. Int. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 289–297.
[12] P. Ekman and W. Friesen, *The Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Santa Clara, CA, USA: Consulting Psychologists Press, 1978.
[13] Y. Tian, T. Kanade, and J. F. Cohn, "Recognizing upper face action units for facial expression analysis," in *Proc. Int. Conf. Comput. Vis. Pattern Recognit.*, 2000, pp. 294–301.
[14] Y. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 97–115, Feb. 2001.
[15] Y. Tong, W. Liao, and Q. Ji, "Facial action unit recognition by exploiting their dynamic and semantic relationships," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 10, pp. 1683–1699, Oct. 2007.
[16] M. S. Bartlett *et al.*, "Fully automatic facial action recognition in spontaneous behavior," in *Proc. 7th Int. Conf. Autom. Face Gesture Recognit.*, Southampton, U.K., 2006, pp. 223–230.
[17] J. F. Cohn, L. Reed, Z. Ambadar, J. Xiao, and T. Moriyama, "Automatic analysis and recognition of brow actions and head motion in spontaneous facial behavior," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, 2004, pp. 610–616.
[18] G. Sandbach, S. Zafeiriou, and M. Pantic, "Binary pattern analysis for 3D facial action unit detection," in *Proc. Brit. Mach. Vis. Conf.*, Surrey, U.K., 2012, pp. 119.1–119.12.
[19] G. Sandbach, S. Zafeiriou, M. Pantic, and D. Rueckert, "Recognition of 3D facial expression dynamics," *Image Vis. Comput.*, vol. 30, no. 10, pp. 762–773, Oct. 2012.
[20] H. Kobayashi and F. Hara, "Facial interaction between animated 3d face robot and human beings," in *Proc. Int. Conf. Syst. Man Cybern.*, 1997, pp. 3732–3737.
[21] G. Gao, K. Jia, and B. Jiang, "An automatic geometric features extracting approach for facial expression recognition based on corner detection," in *Proc. Int. Conf. IEEE Intell. Inf. Hiding Multimedia Signal Process.*, 2015, pp. 302–305.
[22] A. Majumder, L. Behera, and V. K. Subramanian, "Emotion recognition from geometric facial features using self-organizing map," *Pattern Recognit.*, vol. 47, no. 3, pp. 1282–1293, 2014.
[23] Q. Mao, Q. Rao, Y. Yu, and M. Dong, "Hierarchical Bayesian theme models for multipose facial expression recognition," *IEEE Trans. Multimedia*, vol. 19, no. 4, pp. 861–873, Apr. 2017.
[24] M. Huang, Z. Wang, and Z. Ying, "A new method for facial expression recognition based on sparse representation plus LBP," in *Proc. IEEE Int. Congr. Image Signal Process.*, 2010, pp. 1750–1754.
[25] T. Jabid, M. Kabir, and O. Chae, "Robust facial expression recognition based on local directional pattern," *ETRI J.*, vol. 32, pp. 784–794, 2010.
[26] M. Bartlett *et al.*, "Recognizing facial expression: Machine learning and application to spontaneous behavior," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, 2005, pp. 568–573.
[27] A. Dhall, A. Asthana, R. Goecke, and T. Gedeon, "Emotion recognition using PHOG and LPQ features," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recog. Workshops*, 2011, pp. 878–883.
[28] A. Tawari and M. M. Trivedi, "Face expression recognition by cross modal data association," *IEEE Trans. Multimedia*, vol. 15, no. 7, pp. 1543–1552, Nov. 2013.
[29] T. Zhang *et al.*, "A deep neural network-driven feature learning method for multi-view facial expression recognition," *IEEE Trans. Multimedia*, vol. 18, no. 12, pp. 2528–2536, Dec. 2016.
[30] L. Zhong, Q. Liu, P. Yang, J. Huang, and D. N. Metaxas, "Learning multiscale active facial patches for expression analysis," *IEEE Trans. Cybern.*, vol. 45, no. 8, pp. 1499–1510, Aug. 2015.
[31] S. L. Happy and A. Routray, "Robust facial expression classification using shape and appearance features," in *Proc. 8th Int. Conf. Adv. Pattern Recognit.*, 2015, pp. 1–5.
[32] K. Zhao, W.-S. Chu, and H. Zhang, "Deep region and multi-label learning for facial action unit detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3391–3399.
[33] J. Li and E. Lam, "Facial expression recognition using deep neural networks," in *Proc. IEEE Int. Conf. Imag. Syst. Tech.*, Sep 2015, pp. 1–6.
[34] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2016, pp. 1–10.
[35] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via a boosted deep belief network", in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1805–1812.
[36] D. Hamester, P. Barros, and S. Wermter, "Face expression recognition with a 2-channel convolutional neural network," in *Proc. Int. Joint Conf. Neural Netw.*, 2015, pp. 1–8.
[37] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *Proc. Int. Conf. Comput. Vis.*, Dec. 2015, pp. 2982–2991.
[38] M. Xu, W. Cheng, Q. Zhao, L. Ma, and F. Xu, "Facial expression recognition based on transfer learning from deep convolutional networks," in in *Proc. 11th Int. Conf. Natural Comput.*, 2015, pp. 702—708.
[39] A. Babenko and V. Lempitsky, "Aggregating deep convolutional features for image retrieval," in *Proc. Int. Conf. Comput. Vis.*, 2015, pp. 1269–1277.
[40] X. Lu, Z. Lin, X. Shen, R. Mech, and J. Z. Wang, "Deep multi-patch aggregation network for image style aesthetics and quality estimation," in *Proc. IEEE Int. Conf. Comput. Vis.* , 2015, pp. 990–998.
[41] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, and Z. Ambadar, "The extended Cohn-Kanade dataset (CKþ): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 94–101.
[42] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 1998, pp. 200–205.
[43] P. Viola and M. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, May 2004.
[44] L. Zhong *et al.*, "Learning active facial patches for expression analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2562–2569
[45] S. Taheri, Q. Qiang, and R. Chellappa, "Structure-preserving sparse decomposition for facial expression analysis," *IEEE Trans. Image Process.*, vol. 23, no. 8, pp. 3590–3603, Aug. 2014, Art. no. 3590.
[46] M. Liu, S. Li, S. Shan, R. Wang, and X. Chen, "Deeply learning deformable facial action parts model for dynamic expression analysis" in *Proc. Comput. Vis.*, 2014, pp. 1749–1756.
[47] P. Khorrami, T. Paine, and T. Huang, "Do deep neural networks learn facial action units when doing expression recognition?," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2015, pp. 19–27.
[48] K. Subramanian, S. Suresh, and R. Venkatesh Babu, "Meta-cognitive neuro-fuzzy inference system for human emotion recognition," in *Proc. Int. Joint Conf. Neural Netw.*, Jun. 2012, pp. 1–7.
[49] J. A. R. Castillo, A. R. Rivera and O. Chae, "Recognition of face expressions using local principal texture pattern," in *Proc. 19th IEEE Int. Conf. Image Process*, Sep. 2012.
[50] S. H. Lee, K. Plataniotis, and Y. M. Ro, "Intra-class variation reduction using training expression images for sparse representation based facial expression recognition," *IEEE Trans. Affective Comput.*, vol. 5, no. 3, pp. 340–351, Jul.–Sep. 2014.
[51] J. A. R. Castillo, A. R. Rivera, and O. Chae, "Facial expression recognition based on local sign directional pattern," in *Proc. 19th IEEE Int. Conf. Image Process.*, Sep. 2012, pp. 2613–2616.
[52] F. A. M. da Silva, and H. Pedrini, "Effects of cultural characteristics on building an emotion classifier through facial expression analysis," *J. Electron. Imag.,* vol. 24, no. 2, pp. 023015–023015, 2015.



Siyue Xie is currently working toward the Graduate degree at the School of Electronics and Information Engineering, Sun Yat-sen University, Guangzhou, China. His research interests include computer vision and pattern recognition. One particular interest is facial expression recognition.

Haifeng Hu received the Ph.D. degree from Sun Yat-sen University, Guangzhou, China, in 2004. Since July 2009, he has been an Associate Professor with the School of Electronics and Information Engineering, Sun Yat-sen University. He is currently a Visiting Professor with the Robotics Institute of Carnegie Mellon University. He has authored and coauthored more than 80 papers since 2000. His research interests include computer vision, pattern recognition, image processing, and neural computation.