Wireless Personal Communications (2022) 123:3465–3482
https://doi.org/10.1007/s11277-021-09298-y

Sparse Representation Based Facial Expression Classification with Decision-Fusion Based on Compound-Variational Dictionaries

Yan Ouyang[1] **· Peiqi Deng**[2]

Accepted: 20 October 2021 / Published online: 16 November 2021

Abstract
Sparse representation-based classification (SRC) has been successfully used in facial expression recognition, well-known for its high accuracy and robustness to some pollutions such as corruption and occlusion. However, the environmental change and human identity of facial expression image samples still interfere with the performance of the SRC framework and typically yields low accuracy. To improve intra-class variations' robustness, we proposed using compound-variational dictionaries to solve this issue. In addition, a two-stage SRC framework and fusion strategy has also been designed to ensure the dictionaries can achieve better performance. In the process of our work, the first stage depends on the compound dictionary, including the information of apex facial expression image. The second stage relies on a dictionary called variational dictionary constructed by the difference information between expressionless face image and apex facial expression image of one person. Subsequently, the classification results of the two stages are fused by the reconstruction error-based strategy. Due to the existing state-of-the-art SRC techniques concerning the one-stage framework, the experiment results on the Cohn-Kanada (CK+) database indicate that the accuracy of the proposed approach can be improved by about 3%. We also tested the approach on the JAFFE database and achieved nearly 100% accuracy, which verified the generalization performance of our work.

Keywords Region of interest · Local binary patterns · Histogram of oriented gradient · Reconstruction error

1 Introduction

Within the last decades, the sparse representation-based classification (SRC) framework has been widely used in face and facial expression recognition because of its robustness to interferences such as corruptions and occlusions. Wright et al. [1] firstly present the SRC

✉ Yan Ouyang
y_ouyang_o121@yeah.net

[1] Early Warning Academy, Wuhan 430019, China

[2] Hubei Academy of Social Sciences, Wuhan 430020, China

and applied it to face recognition. Their work not only shows the superiority of SRC, but also annouces that feature extraction methods are not necessary to SRC framework when the training face images are sufficient. Their claim is supported by the experimental results that SRC with random projection-based features can outperform some traditional face recognition schemes. However, according to the comparison concerning the experimental results of published works, the feature extraction method impacts the performance of the SRC framework, especially in the FER [2], because effective facial expression feature can stand out specific expression characters better than the raw face images, for instance, Eigenfaces [2, 3], Gabor [2], LBP [4], HOG [5], Deep learning feature [6]. The experimental results also show that the SRC framework performs better than some traditional classification framework, such as Support Vector Machine (SVM) and NearestNeighbour (NN) when noises and blocks pollute facial expression images. In the work of FER based on SRC mentioned above, all the researchers take the overall facial features as the main basis for distinguishing expressions, while ignore the unique features of each type of expressions.

Therefore, with the development of FER research [7, 8], expression feature extraction methods gradually begin to focus on the method of reducing intra-class variation. Before the FER researchers realize the importance of intra-class variation, it already exists in many areas of object detection [9] and interprets as illumination, viewpoint, scale, occlusion, shading, clutter, blur, motion, and imaging noise, etc. In FER work, inra-class variation mainly refers to identity and illumination. In the existing applications of the SRC framework, the main two ways of reducing the impact of intra-class variation are based on dictionary construction and decomposition-reconstruction accuracy. In detail, one way is trying to implement different feature extraction or dictionary optimization method, increasing the difference between sample categories in the dictionary, such as PCA-based dictionary building rely on different images [10], intra-class variation reduction features (IVRF) [7, 8] or KSVD (K-means Singular Value Decomposition) [11]. The other way is trying to optimize the solution process of sparse codings, such as Collaborate Representation Classification (CRC) [12] or extended sparse representation-based classification method [13, 14]. Comparing those two ways, in our opinion, the first way mainly focuses on dealing with the facial expression characteristics. It is the key to improving FER's performance for SRC, while the second way is more universal, which emphasizes proposing a more common and precise recognition framework based on SRC.

In this paper, we conducted the study following the first way, aiming at an exclusive recognition method based on the SRC framework suitable for the FER task and robust to intra-class variation. In practice, reducing intra-class variation can improve the sparsity of the sparse weighted matrix corresponding to the dictionary set up by training samples. Analysis of the existing works indicates two ways to decline the impact of intra-class variation. One way is to extract difference information between different types of facial expression images [7, 8]. While, the other way is to extract difference information between natural face images and specific facial expression images [10, 15, 16]. In our opinion, the requirement of comprehensive training dictionary is much higher in a first way. Because those researchers [7, 8] use the difference information between query images and the IVRF (intra-class variation reduced features) of all seven facial expressions. And this way will lead to the loss of some common features of AUs shared by different kinds of facial expressions. For instance, in the definition of FACS [17] (facial action coding system), the fear expression is composed by AU1 + AU2 + AU4 + AU5 + AU7 + AU20 + AU25 while AU1 + AU2 + AU5 + AU25 + AU26 composes the surprise expression. Those two expressions share four AUs (AU1, AU2, AU5 and AU25), and many characters may be lost when directly subtracted from each other. In other words,

the second way may lead to less information loss than the first way. Recently, some researchers have already realized that fact, for example, Du et al. [15] propose a facial expression method based on the difference expression images between the neutral face and specific expressions, but they did not extract the auxiliary features of neutral face. Lee et al. [18] directly use the difference vector between ICV face image and peak expression image in videos for FER. However, no matter which way we choose, the risk of character information loss can not be removed. Zhe et al. [19] have realized this problem, but they only treat the differential information as an auxiliary information to the full face information.

As a whole, it is noticeable that the overall facial features and differential facial features are both critical to FER. However, they are not treated as equal in many existing works. Therefore, this paper proposes a novel FER method based on compound-variational dictionaries to fuse those two kinds of features equally. Figure 1 indicates the process of the proposed method.

The main contribution of this paper can be summarized as follows:

1. The compound-variational dictionaries are proposed to fuse mixed and difference information while dealing with FER problems. As shown in Fig. 1, the compound dictionary built by facial expression features at the apex is used to reserve some characteristics that may be lost during the difference operation between neutral face and specific expression face. On the other hand, the variational dictionary is used to remove the intra-class variation, such as illumination and identity.
2. A novel FER process based on compound-variational dictionaries and SRC framework is proposed. To achieve better performance, we design a two-stage SRC procedure. Firstly, we reconstruct the mixed information of specific facial expression based on the sparse coefficient of the compound dictionary, and make a preliminary judge for any query image sequences. Secondly, we use the difference information between the neutral face images and face image with specific facial expressions at apex level in one face image sequence to judge different facial expressions with variational dictionaries. Finally, a

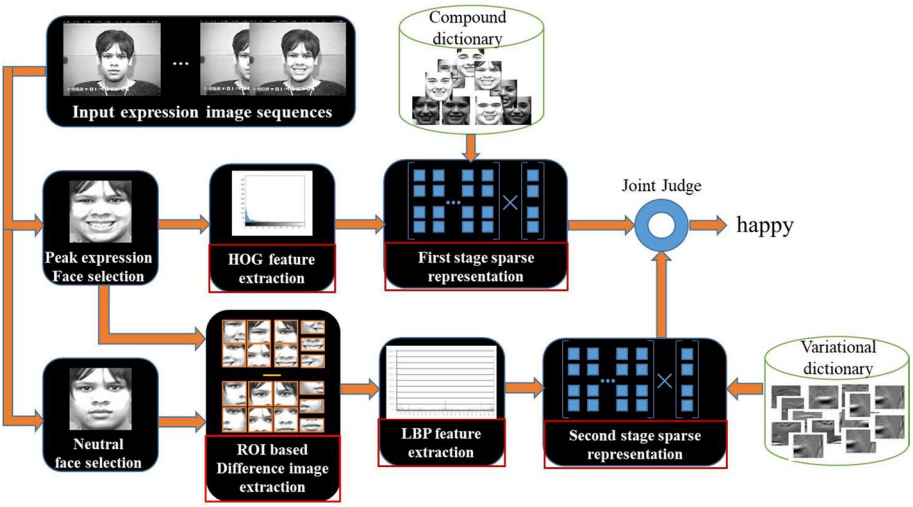


Fig. 1 Flowchart of the proposed method based on decision-fusion and compound-variational dictionaries

decision fusion strategy relies on the reconstruction error is implemented to given the final judgment.

The rest of the paper is organized as follows: Sect. 2 introduces the proposed method for solving the FER task and gives detailed algorithms. Section 3 shows the experimental results and compares the proposed method with existing state-of-the-art methods in FER problems. The conclusion is given in Sect. 4.

2 Methods

2.1 Dictionary Design

In practice, the dictionary is essential for the SRC framework when carrying out the FER task. In early work, many researchers believe that only one sufficient dictionary is enough for the SRC framework to finish the work. However, with further research, many studies found that two or more dictionaries can help the SRC framework to improve its performance under some given situations, such as completing the under-sampled FR challenge [13], simplifying the process of sparse coding [12], and rating the exaggeration of expression [20].

In this paper, inspired by the above work, we suggest using dual dictionaries structure called compound-variational dictionaries to reduce the intra-class variation and eliminate the impact of information loss. The establishing process of dictionaries is shown in Fig. 2.

Generally, it is challenging to align all the essential facial units when dealing with the whole face. So, when design the variational dictionary, we determine some typical Regions of Interest (ROIs) that characterizing the facial action units in face image sequences. As shown in Fig. 3, a representative set of ROIs are selected due to the Action-Unit definition in the FACS system [17] and pre-processed facial feature points. In our study, we extract nine ROIs from each face image.

In practical, firstly, we selected an expressionless face image and defined as d_j in one facial expression sequence. Secondly, we select some facial expression images at the apex

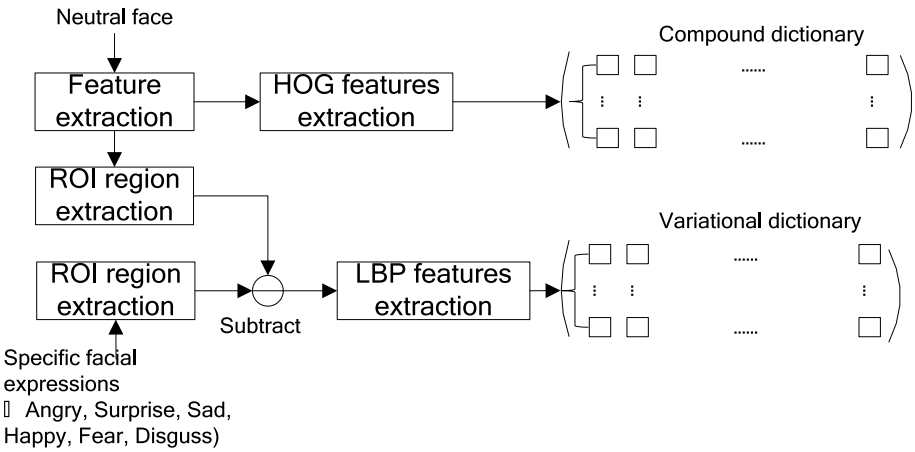


Fig. 2 Establishing process of compound-variational dictionaries

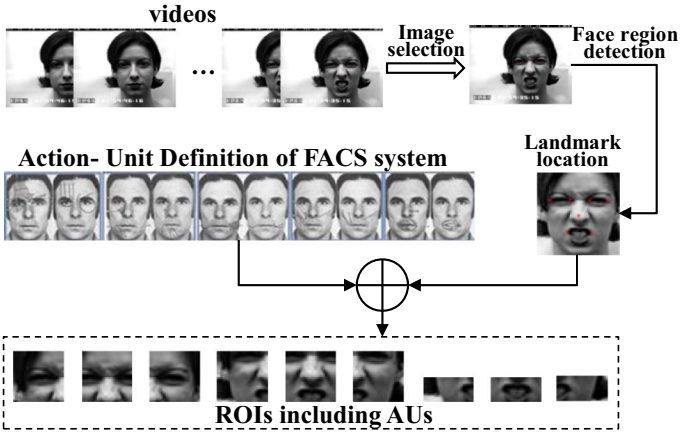


Fig. 3 Illustration of ROIs selection

level and defined as $\left\{b_j^n\right\}$, where n denotes the number of selected frames. The selected ROIs of face image with specific expressions are defined as $\left\{\theta^i(\cdot)\right\}_K$ where $\{i=1, 2, \ldots, 9\}$ denotes the order of ROIs and $\{\mathrm{K}=\mathrm{k}_{\mathrm{sad}}, \mathrm{k}_{\mathrm{surprise}}, \mathrm{k}_{\mathrm{angry}}, \mathrm{k}_{\mathrm{happy}}, \mathrm{k}_{\mathrm{fear}}, \mathrm{k}_{\mathrm{hate}}\}$ denotes the number of training sequences with specific expressions. The compound dictionary D, is formed by merging $\left\{\theta^i\left(\left\{b_j^n\right\}\right)\right\}$ for each person and is arranged into the matrix as below:

$$D \equiv \left[H\left(\left\{\theta^i\left(\left\{b_1^n\right\}\right)\right\}_{k_{\mathrm{sad}}}\right), \ldots, H\left(\left\{\theta^i\left(\left\{b_3^n\right\}\right)\right\}_{k_{\mathrm{angry}}}\right), \ldots, H\left(\left\{\theta^i\left(\left\{b_6^n\right\}\right)\right\}_{k_{\mathrm{hate}}}\right)\right] \in R^{N\times K} \tag{1}$$

where H(·) denotes extracting HOG features and N denotes the feature dimension. The variational dictionary E is formed based on image differencing between d_j and $\left\{b_j^n\right\}$, and arranged into the matrix as below:

$$E \equiv \left[L\left(\left\{\theta^i\left(\left\{b_1^n\right\}-d_1\right)\right\}_{k_{\mathrm{sad}}}\right), \ldots, L\left(\left\{\theta^i\left(\left\{b_3^n\right\}-d_3\right)\right\}_{k_{\mathrm{angry}}}\right), \ldots, L\left(\left\{\theta^i\left(\left\{b_6^n\right\}-d_6\right)\right\}_{k_{\mathrm{hate}}}\right)\right] \in R^{M\times K} \tag{2}$$

where $L(\cdot)$ denotes extracting LBP features from the difference image between $\left\{b_j^n\right\}$ and d_j, M denotes the feature dimension.

2.2 Facial Features Extraction

In many previous work [5, 10, 15], we found that HOG and LBP features have better performance than other features such as Gabor, Haar, Eigenface, Fisherface and Laplacian face, when combined with SRC. So, we use the histogram of oriented gradient (HOG) and to extract the joint information and local binary patterns (LBP) to extract differential information between neutral face image and images with apex facial expressions of one individual. Furthermore, the fineness of feature extraction are also very important [5, 8, 18]. In our previous work [5], we set up some strategies for improving fineness shown as below. When extracting HOG features, the strategies of spatial cell segmentations including: (1) face images are divided by a sliding window(4×4 pixels) with the interval of 2 pixels; (2) face images are

divided by a sliding window (8×8 pixels) with the interval of 4 pixels; (3) face images are divided by a sliding window(16×16 pixels) with the interval of 8 pixels. When extracting LBP features, the strategies including: (1) face images are divided by a sliding window(4×4 pixels) with the interval of 1 pixel; (2) face images are divided by a sliding window(4×4 pixels) with the interval of 2 pixels; (3) face images are divided by a sliding window(8×8 pixels) with the interval of 2 pixels; (4) face images are divided by a sliding window(8×8 pixels) with the interval of 4 pixels. In this issue, we choose all the stragies instead of finding a better one. Because the higher the fineness of features are, the better the recognition accracy will be achieved.

All the extracted HOG features by those three segmentation strategies are union to one compound dictionary. The process is described in Fig. 4. While all the extracted LBP features are union to one variational dictionary.

2.3 Two-Stage SRC (TSSRC) Framework

In this paper, we design a two-stage SRC framework called TSSRC. In the TSSRC framework, each input face image sequence combines two different sub-signals in the compound and variational dictionaries. For a test image sequence y, we first select $p_{neutral}$ and q_{peak} face images, comprehensive characteristics $y_1 = H\left(\theta^i\left(q^{\text{peak}}\right)\right)$ and differential characteristics $y_2 = L\left(\theta^i\left(q^{\text{peak}} - p^{\text{neutral}}\right)\right)$ are extracted. Finally, the optimal sparse coefficient vector $\hat{l}$ and $\hat{m}$ of two-stage SRC are obtained employing the following l_1-norm minimization problem as below:

$$\hat{l} = \arg\min \|l\|_1 \; s.t. \left\|y_1 - Dl\right\|_2 \leq \varepsilon \tag{3}$$

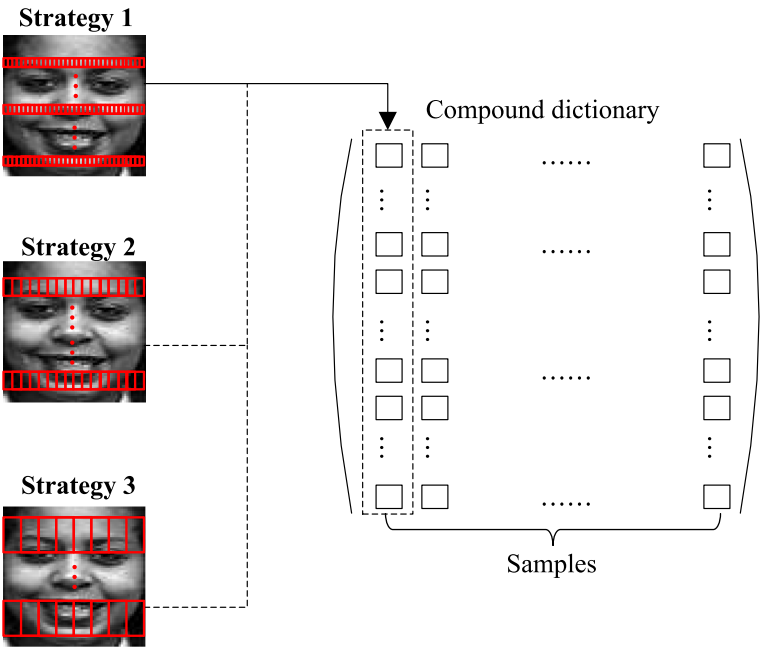


Fig. 4 The building process of compound dictionary

$$\hat{m} = \arg\min \|m\|_1 \quad s.t. \|y_2 - Em\|_2 \leq \varepsilon \tag{4}$$

In physical conception, Eq. (3) means that the input face image is represented by the features of the whole face, Eq. (4) means that the feautres of sepcific facial expressions represent the input face image.

In particular, the preliminary expression class label $T_{j_1}^{\text{pre}}$ and $Y_{j_2}^{\text{pre}}$ are determined by finding the expression class with the maximum of sparse coefficient:

$$T_{j_1}^{\text{pre}} = \arg\max \sum_{j_1=1}^{6} \delta_{j_1}(\hat{l}) \tag{5}$$

$$Y_{j_1}^{\text{pre}} = \arg\max \sum_{j_2=1}^{6} \delta_{j_2}(\hat{m}) \tag{6}$$

where $\delta_{j_1}(\hat{l}) = \left[0, \ldots, l_1^{j_1}, l_2^{j_1}, \ldots\right]$ and $\delta_{j_1}(\hat{m}) = \left[0, \ldots, m_1^{j_1}, m_2^{j_1}, \ldots\right]$ which indicate the sparse coefficient, sub-vectors corresponding to the expression class in $\hat{l}$ or $\hat{m}$.

If $T_{j_1}^{\text{pre}}$ and $Y_{j_2}^{\text{pre}}$ indicate the same expression class, then we can directly output the final classification result. Otherwise, we utilize joint judgment based on the sparse coefficient $\hat{l}$ and $\hat{m}$.

In practical work, we first train two Auxiliary decision dictionaries for D and E by coding each sample with the rest of the samples. Table 1 summarizes the process.

Finally, we use S and Z to integrate the two-stage classification results, as shown in Table 2.

2.4 Facial Expression Recognition Based on TSSRC

In this part, a particular implementation is considered to assess the performance of using TSSRC for FER in facial image sequences as shown in Fig. 5. The main steps of the proposed FER in face image sequences with TSSRC are summarized as follows.

Step 1 Probe expressionless sample y_1 with neutral face and sample y_2^w (w = 3) at apex expression in the probe image sequence are selected. Training samples are used to build compound dictionary D and variational dictionary E.
Step 2 ROI regions of y_1 and y_2^w are found and defined as $\theta^i(y_1)$ and $\theta^i(y_2^w)$.
Step 3 HOG descriptor is applied to sample y_2 to generate feature space $H(y_2^w)$.
Step 4 LBP descriptor is applied to the difference information between $\theta^i(y_1)$ and $\theta^i(y_2^w)$ to generate feature space $L(\theta^i(y_2^w) - \theta^i(y_1))$.
Step 5 SRC is executed for $H(y_2^w)$ and $L(\theta^i(y_2^w) - \theta^i(y_1))$ based on D and E to generate two sparse coefficients vector.
Step 6 If the judgment of two-stage SRC indicates different class labels, we join those sparse coefficients vector based on auxiliary decision dictionaries S and Z to get the final judgment.

Table 1 Training process of auxiliary decision dictionaries

Algorithm 1: Building auxiliary decision dictionaries

Input: compound dictionary D and variational dictionary E

1 for each d_k in D and e_k in E do

2 solve the sparse representation problem to estimate coefficient matrix $\hat{S}_k$ and $\hat{Z}_k$ for d_k and e_k by Eq.8 and Eq.9.

$$\hat{S}_k = \arg\min \|S_k\|_1 \quad s.t. \|d_k - D^{k-1} S_k\|_2 \le \varepsilon \tag{8}$$

$$\hat{Z}_k = \arg\min \|Z_k\|_1 \quad s.t. \|e_k - E^{k-1} Z_k\|_2 \le \varepsilon \tag{9}$$

Where D^{k-1} denotes the dictionary which removes d_k from D and E^{k-1} denotes the dictionary which removes e_k from E.

3 compute the subclass sum in $\hat{S}_k$ and $\hat{Z}_k$ and generate vectors as below:

$$\hat{S}_k^{sum} = \left[\sum \delta_1(\hat{S}_k), \sum \delta_2(\hat{S}_k), \sum \delta_3(\hat{S}_k), \sum \delta_4(\hat{S}_k), \sum \delta_5(\hat{S}_k), \sum \delta_6(\hat{S}_k)\right]$$

$$\hat{Z}_k^{sum} = \left[\sum \delta_1(\hat{Z}_k), \sum \delta_2(\hat{Z}_k), \sum \delta_3(\hat{Z}_k), \sum \delta_4(\hat{Z}_k), \sum \delta_5(\hat{Z}_k), \sum \delta_6(\hat{Z}_k)\right]$$

4 Merge those vectors to build auxiliary decision dictionaries S_{j_1} and Z_{j_2}.

$$S_{j_1} = \left[\hat{S}_{k_{j_1}}^{sum}\right]$$

$$Z_{j_2} = \left[\hat{Z}_{k_{j_2}}^{sum}\right]$$

where $j_1, j_2 = 1,2,3,4,5,6$, k_{j_1} means the sample number of j_1 class in S, k_{j_2} means the sample number of j_2 class in Z.

5 end

2.5 Complexity Analysis

In our method, the main three time-consuming processes are sparse codes by three dictionaries. So, the total computational complexity of our method is $O(m_1 k_1 z_1 + m_2 k_2 z_2 + m_3 k_3 z_3 + m_4 k_4 z_4)$, where z_1,z_2,z_3,z_4 are the number of non-zero entries in the sparse coding results of compound dictionary, variational dictionary and auxiliary decision dictionaries. m_1k_1, m_2k_2, m_3k_3 and m_4k_4 refer to the number of elements in those dictioanries.

3 Experiment Results

To evaluate the performance of our approach, we carry out experiments on the CK + databases [21] and select 310 labelled expression image sequences from 110 subjects. In the CK + database mainly consisting of Western faces, each face image in the image sequences has 68 landmarks, as shown in Fig. 6. In practice, our methods need 7 landmarks to help select ROIs.

Table 2 Reconstruction error based joint judgement

Algorithm 2: Joint judgement of two stage SRC

Input: auxiliary decision dictionaries S and Z, sparse coefficient vector $\hat{l}$ and $\hat{m}$

1 compute the subclass sum in $\hat{l}$ and $\hat{m}$ and generate vectors as below:

$$\rho_{j_1}(\hat{l}) = \left[\sum \delta_1(\hat{l}), \sum \delta_2(\hat{l}), \sum \delta_3(\hat{l}), \sum \delta_4(\hat{l}), \sum \delta_5(\hat{l}), \sum \delta_6(\hat{l})\right]$$

$$\rho_{j_2}(\hat{m}) = \left[\sum \delta_1(\hat{m}), \sum \delta_2(\hat{m}), \sum \delta_3(\hat{m}), \sum \delta_4(\hat{m}), \sum \delta_5(\hat{m}), \sum \delta_6(\hat{m})\right]$$

2 solve the sparse representation problem as shown in Eq. 10 and Eq.11

$$\hat{u} = \arg\min \|u\|_1 \quad s.t. \|\rho_{j_1}(\hat{l}) - S_{j_1} \cdot u\|_2 \le \varepsilon \tag{10}$$

$$\hat{v} = \arg\min \|v\|_1 \quad s.t. \|\rho_{j_1}(\hat{m}) - S_{j_1} \cdot u\|_2 \le \varepsilon \tag{11}$$

3 compute reconstruction error as below:

$$e_{\hat{u}} = \|\rho_{j_1}(\hat{l}) - S_{j_1} \cdot \hat{u}\|_2 \tag{12}$$

$$e_{\hat{v}} = \|\rho_{j_2}(\hat{m}) - Z_{j_2} \cdot \hat{v}\|_2 \tag{13}$$

4 if $e_{\hat{u}} < e_{\hat{v}}$ then $J_{final} = T_{j_1}^{pre}$ else $J_{final} = Y_{j_2}^{pre}$

We selected the first one and the last three frames from each sequence to evaluate our methods in the experiment. Moreover, the performance evaluation of the proposed approach is based on a ten-fold cross-validation and LOSO method. In this study, two eye locations were manually determined, and all cropped face images were rescaled to the size of 64×64 pixels. Figure 7 shows the example face images.

At the beginning, we select one-fold as a testing sample and the rest nine folds as training samples to evaluate the performance of our method, and try to use the experiment result to reveal the relationship between two-stage SRC as shown in Tables 3 and 4.

The result of the first stage SRC shown in Table 3 does better for sad, surprise, happy, and hate expression recognition. While the results of the second stage SRC shown in Table 4 do better for angry and fear expression recognition. Therefore, the two-stage SRC are complementary. Figures 8, 9, 10, 11, 12, 13, 14, 15 and 16 present confusion results of the other nine folds, and the complementarity is not an exception.

The results show that the first stage sparse representation can achieve good performance in most cases. It means that the performance of the methods based on the overall facial features is better than the performance of the methods based on variational facial expression features in most instances. However, in some cases, the recognition accuracy of the first stage sparse representation may appear a sharp decline, for example, the results in Figs. 9 and 11. The decline shows that the recognition results will be affected by the identity. Moreover, that problem cannot be easily overcome when using a one-stage sparse representation method based on overall face features. Thus, the role of the second stage sparse representation is mainly improving the robustness of the first stage sparse representation to some unexpected cases. However, the improvement of robustness will lead to more time-consuming. In the following, we compare the time-consuming of some one-stage sparse representation methods. The time-consuming per image of our method is about one minute and twenty-five seconds. The experiments

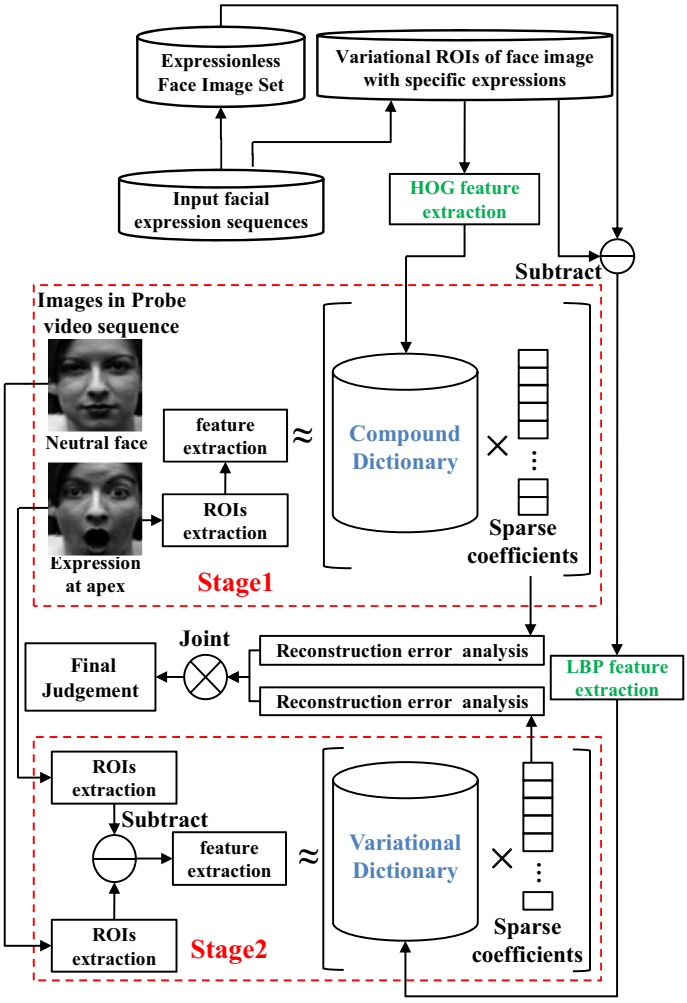


Fig. 5 Block diagram of the proposed approach

are conducted on a notebook with Intel(R) Core(TM) i7-4860HQ and 32 GB RAM. The program is written in MATLAB (Table 5).

For the purpose of fusing the results of two-stage SRC, we propose a novel method based on the analysis of the sparse coefficient vectors. As shown in Figs. 17 and 18, we display two sparse coefficient vectors of different stage SRC when classifying the facial expression of fear in one-fold.

The fourth, fifth, sixth, thirteenth and fifteenth samples may be misclassified in the second stage SRC while they're correctly classifying in the first stage SRC. In this situation, directly using sparsity weights [18] to fuse those two results will face a severe problem, which is the weights are hard to choose by experience. We use the reconstruction error instead of sparsity weights to joint judge those two results based on sparse coefficient vectors.

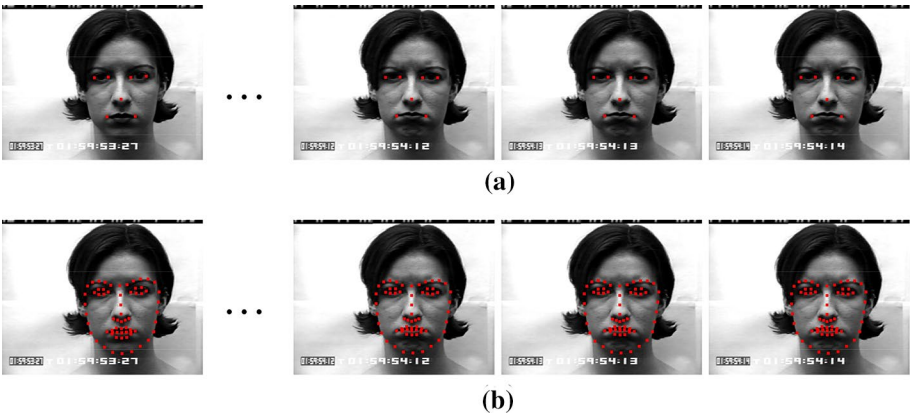


Fig. 6 **a** Image sequence of sad expression in CK + database with 7 landmarks; **b** Image sequence in CK + database with 68 landmarks

Fig. 7 Cropped example face images

Table 3 The confusion result (%) based on one fold in the first stage SRC

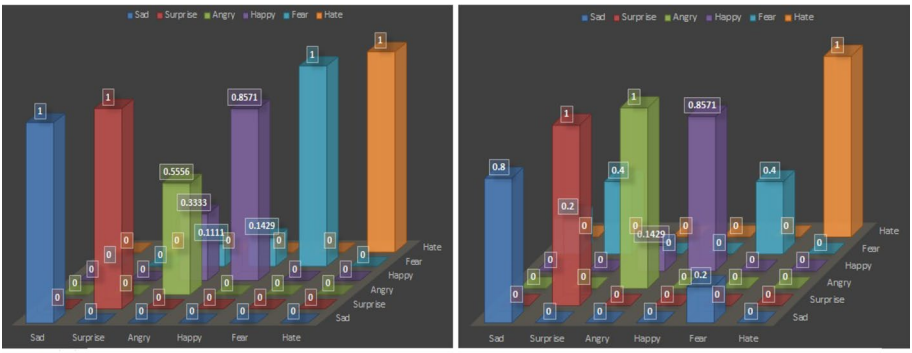
	Sad	Surprise	Angry	Happy	Fear	Hate
Sad	86.67	0	13.33	0	0	0
Surprise	0	100	0	0	0	0
Angry	6.67	0	86.66	0	0	6.67
Happy	0	0	0	100	0	0
Fear	0	0	0	20	80	0
Hate	0	0	0	0	16	84

Table 4 The confusion result (%) based on one fold in the second stage SRC

	Sad	Surprise	Angry	Happy	Fear	Hate
Sad	80	0	20	0	0	0
Surprise	0	91.67	0	0	0	8.33
Angry	6.67	0	93.33	0	0	0
Happy	0	0	7.41	92.59	0	0
Fear	3.7	0	0	3.7	92.6	0
Hate	8.33	0	25	0	0	66.67

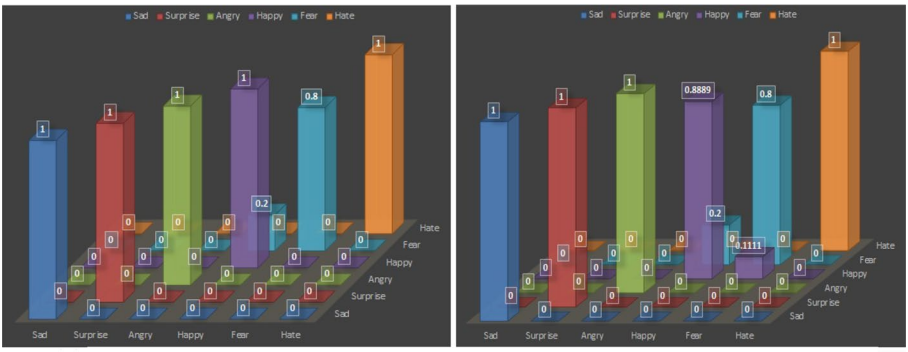
(a) Confusion result in the first stage SRC **(b)** Confusion result in the second stage SRC

Fig. 8 The confusion result based on the first fold



(a) Confusion result in the first stage SRC **(b)** Confusion result in the second stage SRC

Fig. 9 The confusion result based on the second fold

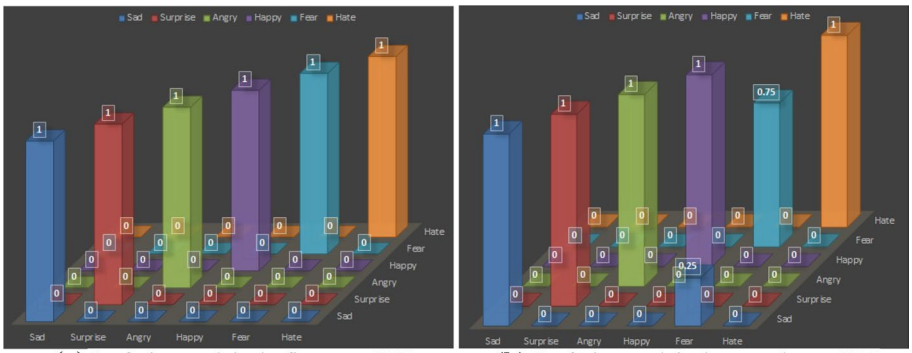


(a) Confusion result in the first stage SRC **(b)** Confusion result in the second stage SRC

Fig. 10 The confusion result based on the third fold

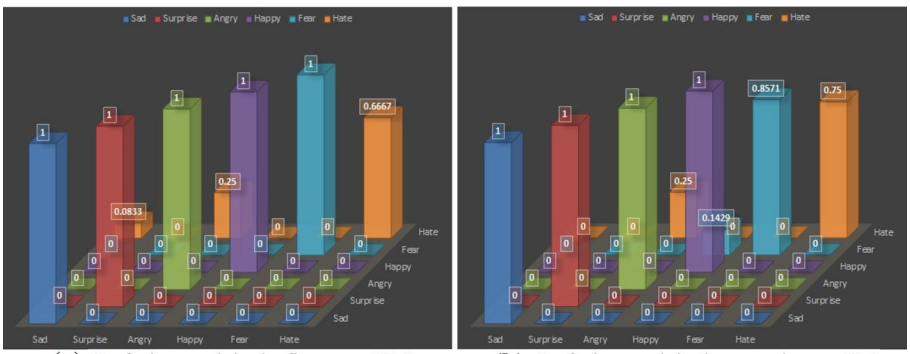
(a) Confusion result in the first stage SRC **(b)** Confusion result in the second stage SRC

Fig. 11 The confusion result based on the fourth fold



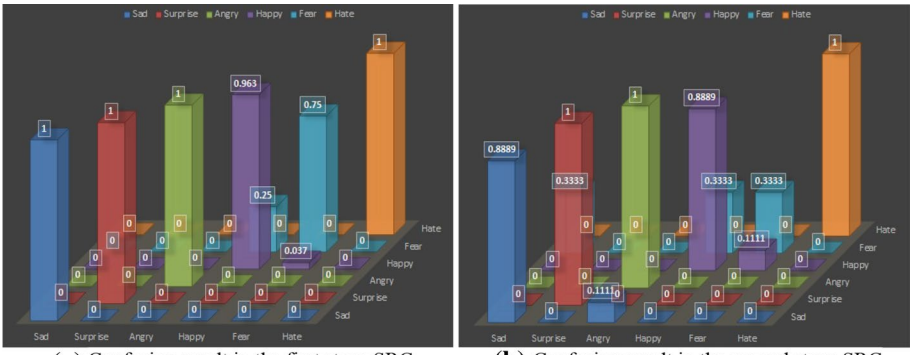
(a) Confusion result in the first stage SRC **(b)** Confusion result in the second stage SRC

Fig. 12 The confusion result based on the fifth fold



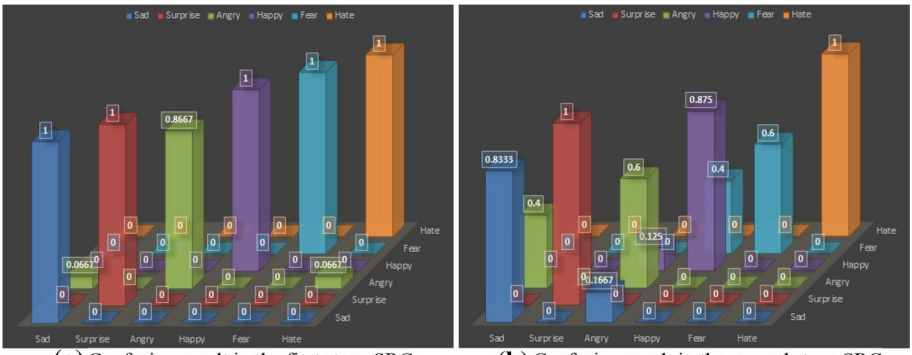
(a) Confusion result in the first stage SRC **(b)** Confusion result in the second stage SRC

Fig. 13 The confusion result based on the sixth fold



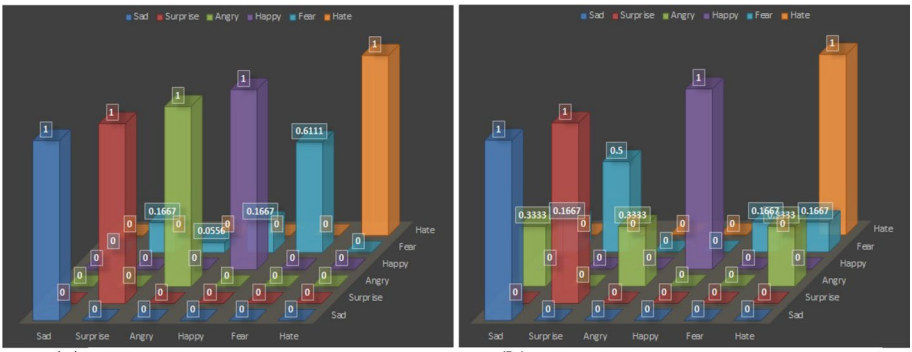
(a) Confusion result in the first stage SRC **(b)** Confusion result in the second stage SRC

Fig. 14 The confusion result based on the seventh fold



(a) Confusion result in the first stage SRC **(b)** Confusion result in the second stage SRC

Fig. 15 The confusion result based on the eighth fold



(a) Confusion result in the first stage SRC **(b)** Confusion result in the second stage SRC

Fig. 16 The confusion result based on the ninth fold

Table 5 The time-consuming of different one-stage sparse representation method

Method	Time-consuming per imge(s)
Eigenface + SRC [1]	1.8493
Laplacianface + SRC [1]	0.8145
Fisherface + SRC [1]	0.0034
HOG + SRC (our)	19
LBP + SRC (our)	60

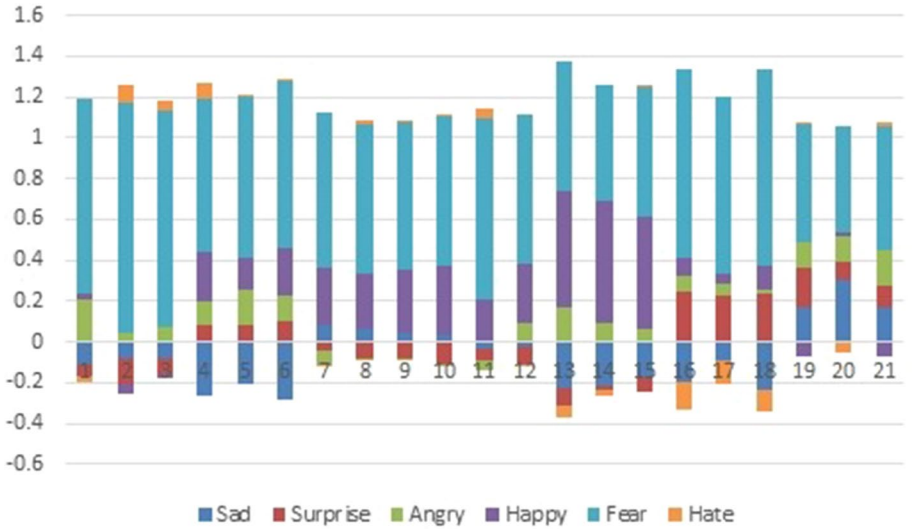


Fig. 17 The sparse coefficient vectors of the first stage SRC

In the following, we compare the final validation result after reconstruction error-based fusion with several classical methods on the CK + database.

Table 6 shows that the proposed TSSRC model achieves the best performance, which indicates that our method has made full use of various facial expression information than the others based on the SRC framework.

We also conduct our method on the JAFFE database, which is mainly consists of Asian faces. The facial expression images are shown in Fig. 19. The results in Table 7 show that the universality of our method is very good.

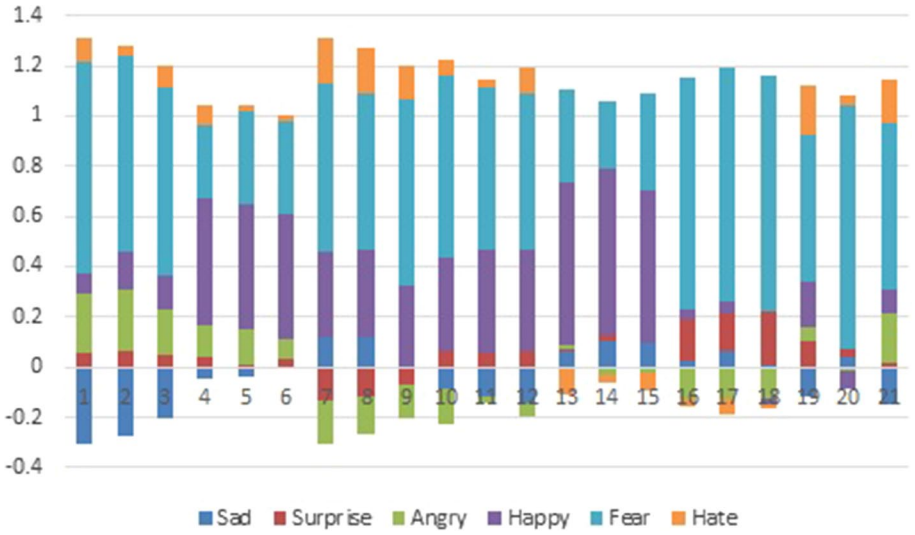


Fig. 18 The sparse coefficient vectors of the second stage SRC

Table 6 FER rates (%) on CK + database

Method	Cross-validation	Results
Our	Tenfold	94.4
Our	LOSO	97.5
Eigenfaces + SRC [2, 3]	Tenfold	79.42
Gabor + SRC [2]	Tenfold	81.52
IVRF + SRC [7]	LOSO	90.5
IVRF + SRC [7]	Tenfold	89.6
IVRF-MRDPL [8]	Tenfold	94.51
ICV + CRC [18]	Tenfold	92.34
ICV + CRC [18]	LOSO	94.9
PCA + SRC [10]	LOSO	97.19
Deep learning feature + SRC [19]	LOSO	85.66

Fig. 19 Image examples in JAFFE database

Table 7 FER rates (%) on JAFFE database

Method	Cross-validation	Results
Our	Tenfold	99.5
Our	LOSO	99.8

4 Conclusions

In this letter, we propose a two-stage SRC to utilize compound and variational expression information fully. Compared with the classical expression recognition methods, our model can obtain better discriminant power. Extensive experiments on the famous datasets (CK+) show the effectiveness of our approach. We also conduct the experiment on the JAFFE database. The accuracy of our method achieves nearly 100%. Comparing those two experimental results based on a different database, we find that our approach can treat the problem of identity very well.

Funding This research was financially supported by the National Natural Science Foundation of China (Grant No: 61503410).

Data Availability The authors confirm that the data supporting the findings of this study are available within the article.

Declarations

Conflicts of interest The authors declare no conflicts of interest.

References

1. Wright, J., Yang, A. Y., Ganesh, A., Sastry, S. S., & Ma, Y. (2009). Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31*(2), 210–227.
2. Cotter, S. F. (2010). Sparse representation for accurate classification of corrupted and occluded facial expressions. In *Proc. ICASSP.* (pp. 838–841).
3. Huang, M. W., & Ming, Z. L. (2010). The performance study of facial expression recognition via sparse representation. In *International conference on machine learning and cybernetics (ICMLC).*
4. Huang, M. W., Wang, Z. W., & Ying, Z. L. (2010). A new method for facial expression recognition based on Sparse representation plus LBP. In *International congress on image and signal processing (CISP)* (pp. 1750–1754).
5. Ouyang, Y., Sang, N., & Huang, R. (2015). Accurate and robust facial expressions recognition by fusing multiple sparse representation based classifiers. *Neurocomputing, 149*, 71–78.
6. Xie, W. C., Jia, X., Shen, L. L., & Yang, M. (2019). Sparse deep feature learning for facial expression recognition. *Pattern Recognition, 96*, 106966.
7. Lee, S. H., & Ro, Y. M. (2014). Intra-class variation reduction using training expression images for sparse representation based facial expression recognition. *IEEE Transactions on Affective Computing, 5*(3), 340–351.
8. Xie, S. Y., Hu, H. F., & Yin, Z. Y. (2018). Facial expression recognition using intra-class variation reduced features and manifold regularization dictionary pair learning. *IET Computer Vision, 12*(4), 458–465.
9. Liu, L., Ouyang, W. L., Wang, X. G., Fieguth, P., Chen, J., Liu, X. W., & Pietikainen, M. (2018). Deep learning for generic object detection: A survey. In: *IJCV*.
10. Mohammadi, M. R., Fatemizadeh, E., & Mahoor, M. H. (2014). PCA-based dicitionary buiding for accurate facial expression recognition via sparse representation. *Journal of Visual Communication and Image Representation, 25*(5), 1082–1092.
11. Aharon, M., Elad, M., & Bruckstein, A. (2006). K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing, 54*(11), 4311–4322.
12. Zhang, L., Yang, M., & Feng, X. (2011). Sparse representation or collaborative representation: Which helps face recognition? In: *ICCV*.
13. Deng, W. H., Hu, J. N., & Guo, J. (2012). Extended SRC: Undersampled face recognition via intra-class variant dictionary. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 34*(9), 1864–1870.

14. Peng, Y. L., Li, L. J., Liu, S. G., Li, J., & Wang, X. L. (2018). Extended sparse representation-based classification method for face recognition. *Machine Vision and Applications, 29*, 991–1007.
15. Du, L. S., & Hu, H. F. (2017). Modified classification and regression tree for facial expression recognition with using difference expression images. *Electronics Letters, 53*(9), 590–592.
16. Kumar, S., Singh, S., & Kuma, J. (2018). Automatic live facial expression detection using genetic algorithm with harr wavelet features and SVM. *Wireless Personal Communications, 103*, 2435–2453.
17. Ekman, P., & Friesen, W. V. (1978). *Facial action coding system* (p. 1). Consulting Psychologists Press.
18. Lee, S. H., Baddar, W. J., & Ro, Y. M. (2016). Collaborative expression representation using peak expression and intra class variation face images for practical subject-independent emotion recognition in videos. *Pattern Recognition, 54*, 52–67.
19. Zhe, S., Raymond, C., & Hu, Z. P. (2018). An extended dictionary representation approach with deep subspace learning for facial expression recognition. *Neurocomputing, 316*(17), 1–9.
20. Ali, M., Karim, F., Hossein, M., & Armon, M. S. (2017). Facial expression recognition using dual dictionary learning. *Journal of Visual Communication and Image Representation, 45*, 20–33.
21. Lucey, P., Cohn, J. F., 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 conf. on computer vision and pattern recognition workshops (CVPRW)* (pp. 94–101). San Francisco, CA, USA.

Yan Ouyang received B.E., M.E. degrees of Software Engineering and Computer Science from Wuhan University of Technology, China, in 2006 and 2009, respectively. He received Ph.D. degree of Control Science and Engineering, from Huazhong University of Science and Technology, Wuhan, China, in 2013. His research interests include facial expression analysis, computer vision and pattern recognition. In recent years, he is mainly working on the application of sparse representation theory in facial expression recognition. In particular, he uses the sparse representation-based classification (SRC) framework to deal with the inherent problems in the process of facial expression recognition.

Peiqi Deng received Ph.D. Degree of Social Policy, from the University of Warwick, Coventry, The United Kingdom, in June 2018. She also received Ph.D. degree of Economics, from Wuhan University, Wuhan, China, in December 2015. She received Master Degree of Comparative and International Social Policy, from the University of York, York, The United Kingdom, in January 2013. Currently, she works as a research assistant in Hubei Academy of Social Sciences, Wuhan, China. Her research interests include Computer Simulation Modeling, Interdisciplinary research, Macro Economics, and Social Research Method, etc.