J. Vis. Commun. Image R. 59 (2019) 176–185

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Video facial emotion recognition based on local enhanced motion history image and CNN-CTSLSTM networks ☆

Min Hu [a,b,*], Haowen Wang [a,b], Xiaohua Wang [a,b], Juan Yang [a], Ronggui Wang [a]

[a] School of Computer and Information of Hefei University of Technology, China
[b] Anhui Province Key Laboratory of Affective Computing and Advanced Intelligent Machine, 230009 Hefei, China

ARTICLE INFO

Article history:
Received 8 June 2018
Revised 10 December 2018
Accepted 21 December 2018
Available online 23 December 2018

Keywords:
Video emotion recognition
Motion history image
LSTM
Facial landmarks

ABSTRACT

This paper focuses on the issue of recognition of facial emotion expressions in video sequences and proposes an integrated framework of two networks: a local network, and a global network, which are based on local enhanced motion history image (LEMHI) and CNN-LSTM cascaded networks respectively. In the local network, frames from unrecognized video are aggregated into a single frame by a novel method, LEMHI. This approach improves MHI by using detected human facial landmarks as attention areas to boost local value in difference image calculation, so that the action of crucial facial unit can be captured effectively. Then this single frame will be fed into a CNN network for prediction. On the other hand, an improved CNN-LSTM model is used as a global feature extractor and classifier for video facial emotion recognition in the global network. Finally, a random search weighted summation strategy is conducted as late-fusion fashion to final predication. Our work also offers an insight into networks and visible feature maps from each layer of CNN to decipher which portions of the face influence the networks' predictions. Experiments on the AFEW, CK+ and MMI datasets using subject-independent validation scheme demonstrate that the integrated framework of two networks achieves a better performance than using individual network separately. Compared with state-of-the-arts methods, the proposed framework demonstrates a superior performance.

1. Introduction

Facial analysis has been a long-standing research topic in the computer vison, some of its application such as face recognition, facial expression recognition has attracted much attention recent years owing to advanced network architectures. As for face recognition, it has been the prominent biometric technique for identity authentication and has been widely used in many areas. As for facial expression recognition, real-time automated analysis of facial expression in video plays a crucial role in developing a human-robots interaction interface.

Three modules are both typically needed for such two aforementioned tasks, a face detector, a fiducial point detector and a feature descriptor. Face detector should be capable of detecting faces with varying pose, illumination, and scale, fiducial points are used to align the faces to normalized canonical coordinates to mitigate the effects of in-plane rotation and scaling [1]. Common face detector and fiducial point detector can be used to solve FR and FER problem while specific feature descriptor is needed to handle these two tasks. For face recognition, an ideal face features are expected to have smaller maximal intra-class distance than minimal inter-class distance [2]. That is to say, faces of same person should have smaller distance than faces come from different person in desired features. When it comes to FER, we expect a feature that can achieve smaller distance of same expressions no matter which people these expressions are from.

The key point of solution to the problem is how to find a proper descriptor. In the early 2000s, local-feature-based descriptors, such as Gabor [3] and LBP [4], are proved to achieve robust performance through some invariant properties of local filtering. LBP, as well as its variants [5–7], these binary descriptors are proposed due to their efficiency comparing with those high-dimensional real-valued descriptors. Since 2010, learnable local descriptors [8–10] were introduced to represent the face, in which local filters are learned for better distinctiveness, and the encoding codebook is learned for better compactness.

Although shallow representations aforementioned achieve impressive performance, deep learning methods use a cascade of multiple layers of processing units for feature extraction proved

☆ This paper has been recommended for acceptance by Zicheng Liu.

* Corresponding author.
E-mail addresses: jsjxhumin@hfut.edu.cn (M. Hu), 2016170718@mail.hfut.edu.cn (H. Wang), xh_wang@hfut.edu.cn (X. Wang).

https://doi.org/10.1016/j.jvcir.2018.12.039
1047-3203/

to improve the capabilities of machines in understanding faces [11]. Convolutional Neural Network [12] was developed to ease feature selection and achieve better results than the machine learning methods that already existed. Many researchers have developed similar methods [13–16] based on CNN for human expression recognition. Traditional practices commonly use the whole aligned face of RGB images as the input of the network to learn features for FER. However, Sepidehsadat Hosseini et al. [17] have shown that finding an appropriate feature for the given problem may still be an important issue to tackle since they can enhance the performance of CNN-based algorithms. Encouraged by this conclusion, we aim to find a hand-craft feature to improve the performance of CNN.

Unlike face recognition, expression is a dynamic process. Compared to single-image recognition, the temporal correlations between expression frames of a video provide additional motion information for video recognition. However, it has been proved that MHI (Motion-History Images) [18] could effectively extract dynamic texture to address the problem of facial expression recognition. MHI was firstly proposed to detect human movement and one of its advantages is that a range of times may be encoded in a single frame. Despite its advantages, some limitations are discovered in MHI. MHI may preserve the dominant movement information which is unrelated to emotion expression, while ignoring some subtle movements that probably effectively distinguish the movements of crucial face components (e.g., opening of mouth and rising of eyebrows).

Since facial landmarks are natural locator of facial components, many researchers use facial landmarks to enhance facial component in different tasks [19–21]. Behzad Hasani et al. [19] incorporate facial landmarks in their 3D CNN for emotion recognition by replacing the shortcut in residual unit of 3D CNN with element-wise multiplication of facial landmarks and the input tensor of the residual unit, showing leading performance on several emotion datasets. However, this method is computationally expensive, due to its complex network structure, and numerous parameters. Our method addresses the limitations of MHI and enhances 3D CNN mentioned before by an element-wise multiplication of facial landmarks and difference image to update MHI template, which generates attention-aware dynamic features that enable more distinct representations of subtle movement of crucial facial parts.

In this paper, we proposed a novel method named LEMHI (Local Enhanced Motion History Image), which extracts temporal relations of consecutive frames in a video sequence using MHI, with facial landmarks used to emphasize on more expressive facial components. Furthermore, we followed the intuition of temporal segment LSTM by Chih-Yao Ma et al. [22] and utilized VGG networks and temporal segment LSTM to classify video human expression. Then a random search [23] weighted summation strategy is selected as our late-fusion fashion to combine each predication scores of two models into the final score.

The rest of the paper is organized as follows: Section 2 provides an overview of the related work. Section 3 explains our methods proposed in this research. Experimental results and their analysis are presented in Section 4 and finally the paper is concluded in Section 5.

2. Related work

2.1. Motion history image

The construction of a binary motion energy image (MEI) or binary motion region (BMR) was firstly presented by Davis [24], which represented where motion occurred in an image sequence. The MEI describe the motion-shape and spatial distribution of a motion while an MHI is generated based on MEI. MHI H(x, y, t) could be computed from an update function $\psi(x,y,t)$:

$$\mathrm{H}(x,y,t)=\begin{cases}\tau & if \quad \psi(x,y,t)=1\\ \max(0,H_\tau(x,y,t-1)-\delta) & otherwise\end{cases} \tag{1}$$

Here, (x, y) and t show the position and time, $\psi(x,y,t)$ signal object's presence (or motion) in the current video image, the duration τ decides the temporal extent of the movement, and δ is the decay parameter. This update function is called for every new video frame analyzed in the sequence. Usually, the MHI is generated from a binarized image, obtained from frame subtraction, using a threshold ξ:

$$\psi(x,\mathrm{y},t)=\begin{cases}1 & if \quad D(x,y,t)\geqslant\xi\\ 0 & otherwise\end{cases} \tag{2}$$

where $D(x,y,t)$ is defined with difference distance Δ as:

$$D(x,y,t)=|B(\mathrm{x},\mathrm{y},t)-B(x,y,t\pm\Delta)| \tag{3}$$

Here, $B(\mathrm{x},\mathrm{y},t)$ is the intensity value of pixel location with coordinate (x, y) at the t frame of the image sequence.

2.2. Long short term memory network

For general-purpose sequence modeling, LSTM as a special RNN structure has proven stable and powerful for modeling long-range dependencies in various previous studies [21,22]. An LSTM network computes a mapping from an input sequence $\mathrm{x}=(x_1,x_2,\cdots,x_t)$ to an output sequence $\mathrm{y}=(y_1,y_2,\cdots,y_t)$ by calculating the network unit activations using the following equations iteratively from t = 1 to t:

$$i_t=\sigma(W_{ix}x_t+W_{im}m_{t-1}+W_{ic}\circ c_{t-1}+b_i) \tag{4}$$

$$f_t=\sigma(W_{fx}x_t+W_{fm}m_{t-1}+W_{fc}\circ c_{t-1}+b_f) \tag{5}$$

$$c_t=f_t\circ c_{t-1}+i_t\circ\tanh(W_{cx}x_t+W_{cm}m_{t-1}+b_c) \tag{6}$$

$$o_t=\sigma(W_{ox}x_t+W_{om}m_{t-1}+W_{oc}\circ c_t+b_o) \tag{7}$$

$$m_t=o_t\circ\tanh(c_t) \tag{8}$$

$$y_t=W_{ym}m_t+b_y \tag{9}$$

where the W terms denote weight matrices, the b terms denote bias vectors, σ is the logistic sigmoid function, i, f, o and c are the input gate, forget gate, output gate and cell activation vectors respectively, all of which are the same size as the cell output activation vector m, '$\circ$' is the element-wise product of the vectors, g and h are the cell input and cell output activation functions, generally tanh.

3. Proposed method

3.1. Local enhanced motion history image

Although the motion history image has achieved fruitful results in the field of human action recognition, it is still difficult to capture these subtle motions of facial component. Meanwhile, there is usually background noise when MHI is generated.

Fortunately, as oppose to general object recognition task, in FER, our approach has the advantage of extracting facial landmarks and using this information to improve the recognition rate.

Naturally, we use facial landmarks to improve traditional MHI, so that the major facial components (such as eyebrows, lip corners, eyes, etc.) and other parts of the face that function less in expressing facial expressions could be differentiated and the background noise could be controlled. The concrete practice is as follows:

Firstly, the difference image $D(x,y,t)$ is obtained from frame subtraction, then attention-aware mask of facial landmarks

$M(x,y,t)$ are generated by detected facial landmarks. $M(x,y,t)$ is defined as follows:

$$\mathrm{M}(x,y,t) = \begin{cases} \alpha & (x,y) \in L \\ \alpha - 0.1\alpha d_{M(L,P)} & (x,y) \in W \\ 1 & otherwise \end{cases} \tag{10}$$

Here L is a set of landmarks coordinates while W represents a set of coordinates of pixels that surround landmarks. Weight α can be seen as a weight assigned to every pixels of mask, and it is obvious that landmarks have the highest weights α while their surrounding pixels have lower weights proportional to their distance from the corresponding facial landmark. We choose Manhattan distance $d_{M(L,P)}$ as mentioned distance with a linear weight function. An element-wise multiplication of the mask $M(x,y,t)$ and difference image $D(x,y,t)$ is defined as follows:

$$E(x,y,t) = M(x,y,t) \circ D(x,y,t) \tag{11}$$

where $E(x,y,t)$E(x, y, t) represents enhanced difference image, '$\circ$' is Hadamard product symbol. A threshold ξ is calculated to binarize enhanced difference image:

$$\psi(x,y,t) = \begin{cases} 1 & if \quad E(x,y,t) \geqslant \xi \\ 0 & otherwise \end{cases} \tag{12}$$

Like the traditional MHI, we utilize $\psi(x,y,t)$to update MHI template as follows:

$$H(x,y,t) = \begin{cases} \tau & if \quad \psi(x,y,t) = 1 \\ \max(0, H_\tau(x,y,t-1) - \delta) & otherwise \end{cases} \tag{13}$$

The threshold ξ is selected by iterative method, and the full algorithm for iterative method is presented in Algorithm 1.

Algorithm 1: iterative method to calculate ξ.

Input: Input image I_0, gray value ensemble V, value mean Z_{mean}
Variable: foreground F, background B, temp ξ, T0
Output: threshold ξ
$T_0 = Z_{mean}$, T = 0
While T_0 != T **do**:
  for z in V **do**:
    if z <= Z_{mean}:
      add z to ensemble F
    else:
      add z to ensemble B
  end
  for f in ensemble F **do**:
    calculate mean f_{mean}
  end
  for b in ensemble B **do**:
    calculate mean b_{mean}
  end
  T_0 = T
  T = (f_{mean} + b_{mean})/2
ξ = T
Output the ξ

Algorithm 1. The iterative method to calculate ξ. Result T is the final ξ.

Fig. 1 shows the contrast of MHI and LEMHI with an expression from CK+. The landmarks window mentioned before refers to the pixel area around each landmark, and the size of the window will affect the capturing of facial movement. As shown in Fig. 2, it is hard for a smaller window to completely cover the areas of main facial component. Larger windows mean that there may be overlaps between different windows. Furthermore, flexible enhancement factor can also lead to diverse recognition rate. The influence of enhancement factor α and landmarks window on recognition rate will be investigated in the following experimental part.

3.2. Cross temporal segment LSTM

LEMHI pay more attention to important facial components while ignoring motions of inferior part of face such as cheeks, which results in inadequate exploiting features. To address this problem, we train a Cross Temporal Segment LSTM with CNN features exploited from every frames. Here CNN plays a role in spatial features exploiting with Temporal Segment LSTM taking up extracting temporal features. In a similar approach, Chih-Yao Ma et al. [19] proposed Temporal Segment LSTM to recognize human actions and achieve state of art performances. We improved this network and successfully applied to video motion recognition. Furthermore, we investigated the influence of segment manner to recognition rate. The structure of our Cross Temporal Segment LSTM is shown in Fig. 3. Features generated by CNN are sequentially segmented with one feature overlapped to fed into LSTM.

Formally, given a video V with t frames $\{f_1, f_2, \cdots, f_t\}$, we divide $\{f_1, f_2, \cdots, f_t\}$ into K segments $\{T_1, T_2, \cdots, T_k\}$ of equal durations. Then the spatial-temporal network models can be represented as follows:

$$N(f_1, f_2, \cdots, f_t) = F_{soft\max}(F_{fc}(h_k)) \tag{14}$$

$$h_k = R(h_{k-1}, p_k; w_r) \tag{15}$$

$$p_k = p(F_{cov}(f_i; w_c), \cdots, F_{cov}(f_j; w_c))\, f_i, \cdots, f_j \in T_k \tag{16}$$

F_{cov} is the function representing a Convolution operation with parameters W_c which operates on every frame in snippet T_k and return CNN features, p is a pool operation. R is update function in LSTM, P_k is input of this function while h_{k-1} is hidden state of LSTM, the outputs h_k connected with a fully connection layer and softmax function for predicting the probability of each expression class for the whole video. Meanwhile, we choose the widely used standard categorical cross-entropy loss as final loss and stochastic gradient descent (SGD) to learn the model parameters.

3.3. Integrated framework of LEMHI-CNN and CNN-RNN

Previous research suggested that assemblies of multiple networks can outperform an individual network [39]. Fan et al. [13] have proved that the prediction results of weighted fusion CNN-RNN classifier and 3D convolutional neural network effectively improve the recognition rate of facial expression. Considering the complementarity between LEMHI-CNN model and the CNN-RNN model, we combined the prediction results of the two models into one results. The overview of the integrated framework is shown in Fig. 4. This system is divided into two models. In the first model, video RGB image sequence is used as the input, and the VGG-16 is treated as a feature extractor. VGG16 model contains 13 convolution layers, 5 pool layers, 3 full connection layers and one softmax layer. All of the network configurations are listed in Table 1.

Different layer of VGG network have different capacity to encode features, we empirically evaluated several layers of VGG network for feature exploiting. Among all of these layers, FC-6 is chosen to report our final recognition accuracies.

The second model takes the LEMHI feature of video frame sequence as input, using conventional VGG16 as feature classifier to get predication results. For the results obtained from these two models, we adopt the weighted fusion method. Assuming that

Fig. 1. Contrast of MHI template and LEMHI template. It shows that local enhanced MHI with facial landmarks facilitate capturing subtle motions in human face and the background noises are restrained effectively.

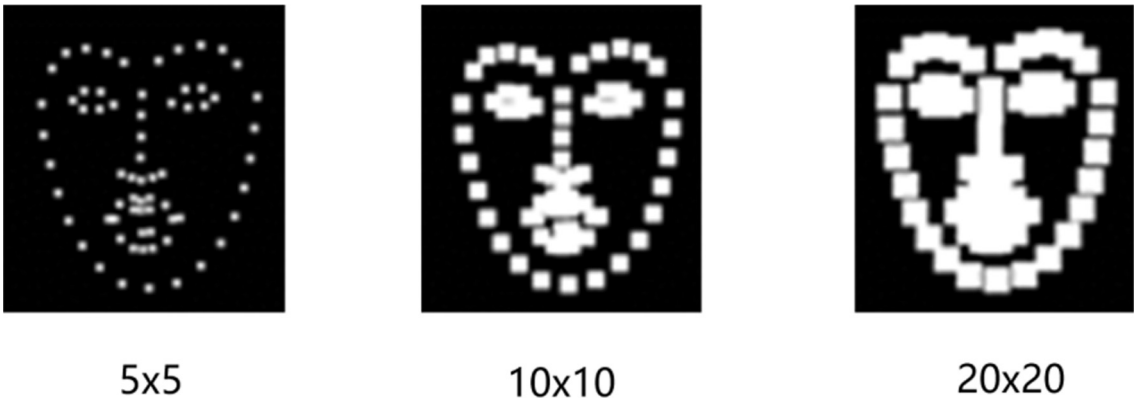


Fig. 2. Masks with diverse landmark windows.

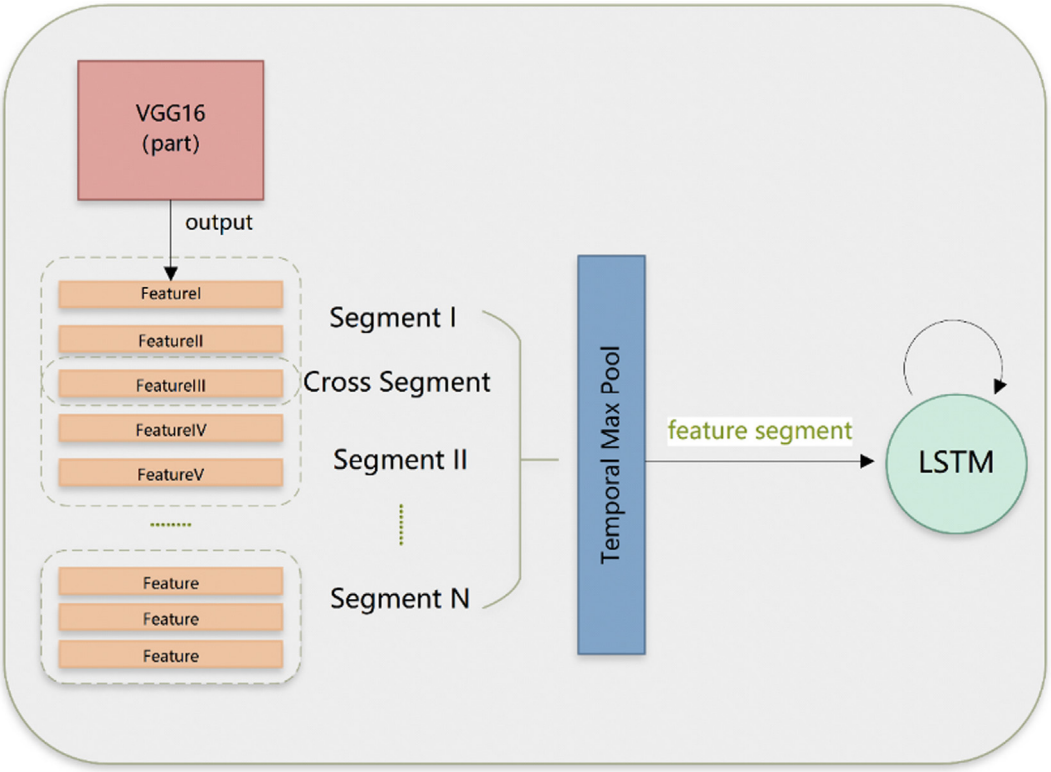


Fig. 3. Cross temporal segment LSTM architecture.

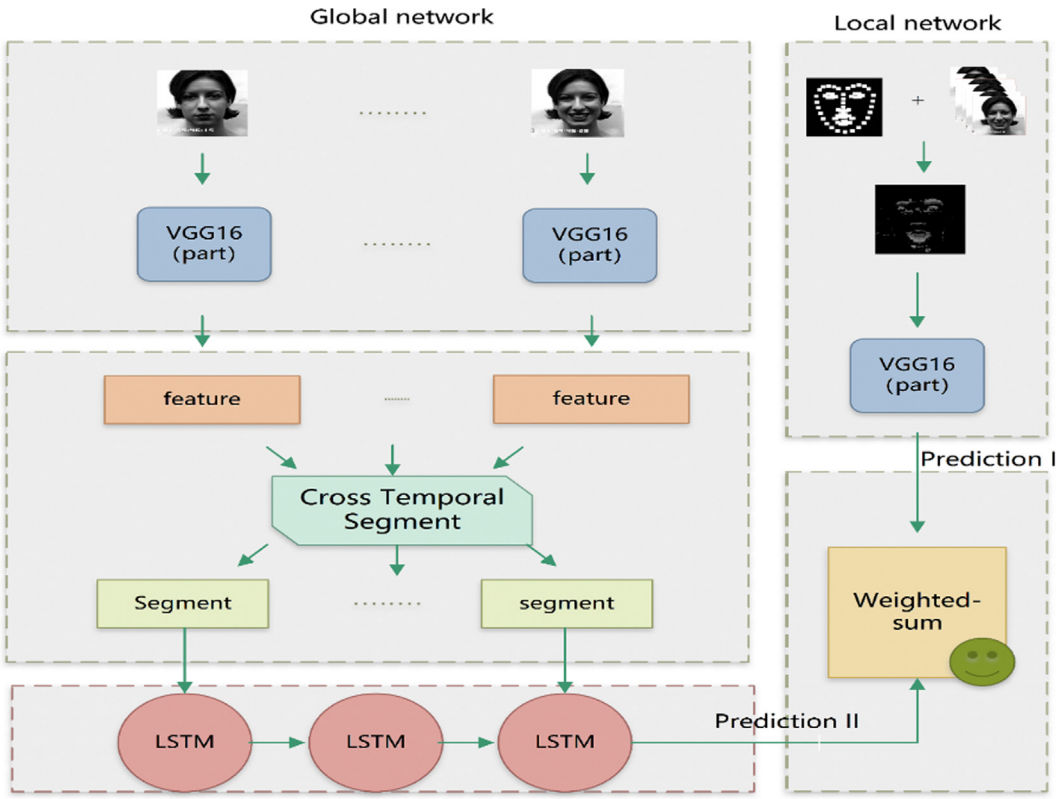


Fig. 4. The overview of the integrated framework.

Table 1
Configurations of all the networks.

	VGG(Part V)	VGG + CTSLSTM
Pre-train	Fer2013	Fer2013
	cov3-64	cov3-64
	maxpool/2	maxpool/2
	cov3-128	cov3-128
	cov3-128	cov3-128
	maxpool/2	maxpool/2
	cov3-256	cov3-256
	cov3-256	cov3-256
	cov3-256	cov3-256
	maxpool/2	maxpool/2
	cov3-512	cov3-512
	cov3-512	cov3-512
	cov3-512	cov3-512
	maxpool/2	maxpool/2
	cov3-512	cov3-512
	cov3-512	cov3-512
	cov3-512	cov3-512
	maxpool/2	maxpool/2
	FC4096	FC4096
	FC4096	Temporal maxpool Lstm-128
Output	FC7/FC6	FC7/FC6

prediction vector obtained from CNN-RNN is P while prediction vector got from LEMHI-CNN is Q, then the final prediction can be defined as follows:

$$R = \lambda P + (1-\lambda)Q \quad 0 \leqslant \lambda \leqslant 1 \tag{17}$$

R represents final prediction result. Here weight λ is determined by random search method. Weights λ are sampled uniformly from [0.0, 1.0] followed by per class re-scaling, so that they sum up to 1. Then the best sampled weights are chosen based on the validation performance. After an initial random search with 100,000 iterations, we perform a local random search around the best set of weights found so far. This local random search consists of sampling weights from a Gaussian with mean set to the current best set of weights and standard deviation σ of 0.5. The current best λ is updated as soon as a new best is found. After every 100,000 iterations, the σ is decreased by a factor of 0.9 and the local search is stopped when σ is smaller than 0.0001.

4. Experiments and results

In this section, we briefly review the databases we used for evaluating our method. We then report the results of our experiments using these databases and compare the results with the state of the arts method.

4.1. Face expression databases

Since our method is designed for video emotion recognition, databases that contain only independent unrelated still images of facial expressions such as **SFEW** [25], **FER2013** [26] cannot be training or testing data for our method. We evaluate our proposed method on **MMI** [27], extended **CK+** [28], **AFEW** [29] which contain videos of annotated facial expressions. Video sequences in MMI and CK+ database are taken in laboratory, while AFEW contains expression video in wild. We will give a briefly review to contents of these databases as follows:

MMI: The MMI database contains more than 20 subjects, whose age ranging from 19 to 62, with different ethnicities (European, Asian, or South American). In MMI, the subjects' facial expression start from the neutral state to the apex of one of the six basic facial expressions and then returns to the neutral state again. Subjects

Table 2
Relationship between number of training sequences and accuracy.

	Angry	Disgust	Fear	Happy	Neural	Sad	Surprise
Sequences	133	74	81	150	144	117	74
Accuracy (%)	69	17	22	72	58	56	25

Table 3
Relationship between number of training sequences and accuracy after extending training data.

	Angry	Disgust	Fear	Happy	Neural	Sad	Surprise
Sequences	133	135(74 + 61)	121(81 + 40)	150	144	117	124(74 + 50)
Accuracy (%)	63	41	38	74	55	52	35

Fig. 5. Sequences fail to be facial landmarks located.

were instructed to display 79 series of facial expressions, six of which are prototypic emotions (angry, disgust, fear, happy, sad, and surprise). We extracted static frames from each sequence. Afterwards, we divided videos into sequences of sixteen frames to shape the input tensor for our models.

CK+: The extended Cohn-Kanade database (CK+) contains 593 videos from 123 subjects. However, only 327 sequences from 118 subjects contain facial expression labels. Sequences in this database start from the neutral state and end at the apex of one of the six basic expressions (angry, contempt, disgust, fear, happy, sad, and surprise). CK+ primarily contains frontal face poses only. In order to make the database compatible with our network, we pick up the last sixteen frames of each sequence as an input sequence in our models.

AFEW: The AFEW database is a series of fragments from 75 movies and 330 subjects, whose age ranging from 1 to 70. It contains 7 expressions, including anger, disgust, fear, happy, neutral, sadness and surprise. There are 773 movie segments for training and 383 movie segments for validation. However, the test set of AFEW is not considered as our training data for the reason that it is not labeled.

4.2. Pre-processing of video frames

If all the original video frames are taken as input directly, especially when frames come from AFEW, a poor distinctive ability is obtained for all kinds of emotions, with an average accuracy no more than 20%. Thus it is necessary to detect face from frames before we fed them into our models. There are three steps to pre-process these original frames: (1) **Face detect**. (2) **Face align**. (3) **Input normalization**. After pre-processing, we get sequences of 224 × 224 size face image as our input tensor. Altogether there are 16 frames in each sequence.

In the experiment, we find that highly imbalance of different expression will result in low accuracy of corresponding expression, which means that the less samples of specific expression are, the lower accuracy this expression will achieve. When CNN-RNN model is trained with AFEW dataset, the proportion of "disgusting" expression to whole expression is far less than 1/7, which leads to a low rate of "disgusting" expression recognition. As shown in Table 2, sequence represents the number of training data for each expression, and the accuracy is evaluated on the validation set by our CNN-RNN model.

A measure was taken to alleviate the imbalance problem. We clipped several videos of three expressions from movies by ourselves, including 61 segments of "disgusting", 40 "fear" and 50 "surprise" facial expression videos, as extended training data for our models.

From the comparison of Tables 2 and 3, we can see that after extending the training data, the accuracy of extended expression has been significantly improved.

We should mention that this extending training data is only used as a pre-train data for our CNN-RNN model. Then the pre-trained model will be fine-tuned by training data from other datasets.

Table 4
Accuracy of LEMHI using several combinations of different landmark window and α on classification of seven facial expressions on CK+ dataset with subject-independent validation.

	30 × 30	20 × 20	10 × 10	5 × 5
$\alpha = 1.5$	80.84	79.36	77.32	71.84
$\alpha = 2.0$	81.21	**83.22**	78.39	71.31
$\alpha = 2.5$	79.49	81.30	77.94	69.64
$\alpha = 3.0$	72.63	77.34	70.26	67.32

4.3. Experimental results

A subject-independent scheme is used to evaluate the performance of the proposed framework with each database split into training and validation sets in a strict subject independent manner. In all the experiment we report the results using 5-fold cross-validation technique and then averaging the recognition rates over five folds.

According to AFEW database, there are several sequences failed to be located facial landmarks due to facial occlusion and large angle of head rotation, as is shown in the Fig. 5. All of these sequences are not considered as our training or test data for LEMHI-CNN models.

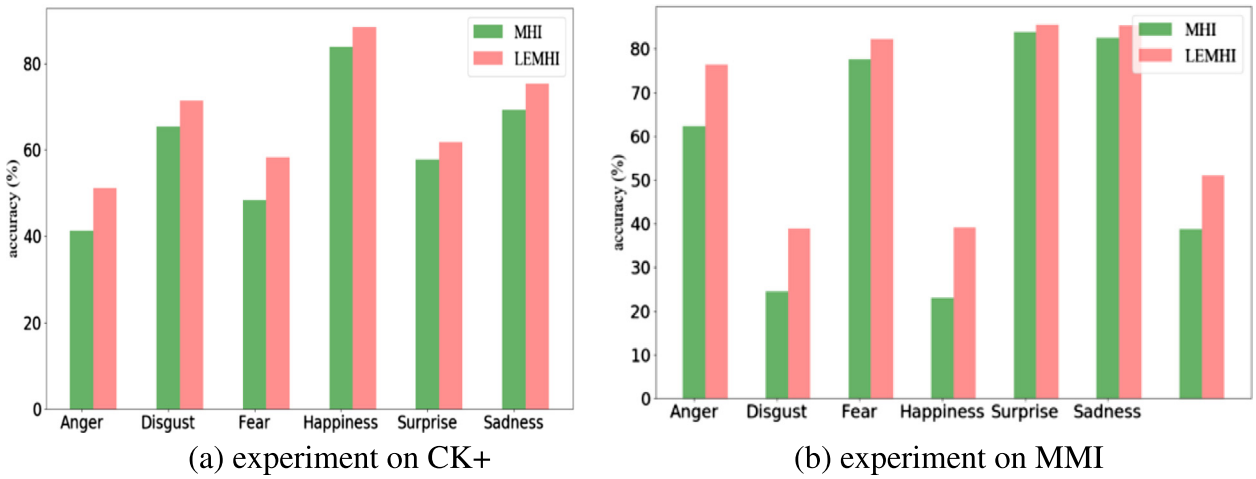


Fig. 6. Recognition rates of all expressions using MHI and LEMHI on CK+ and MMI datasets.

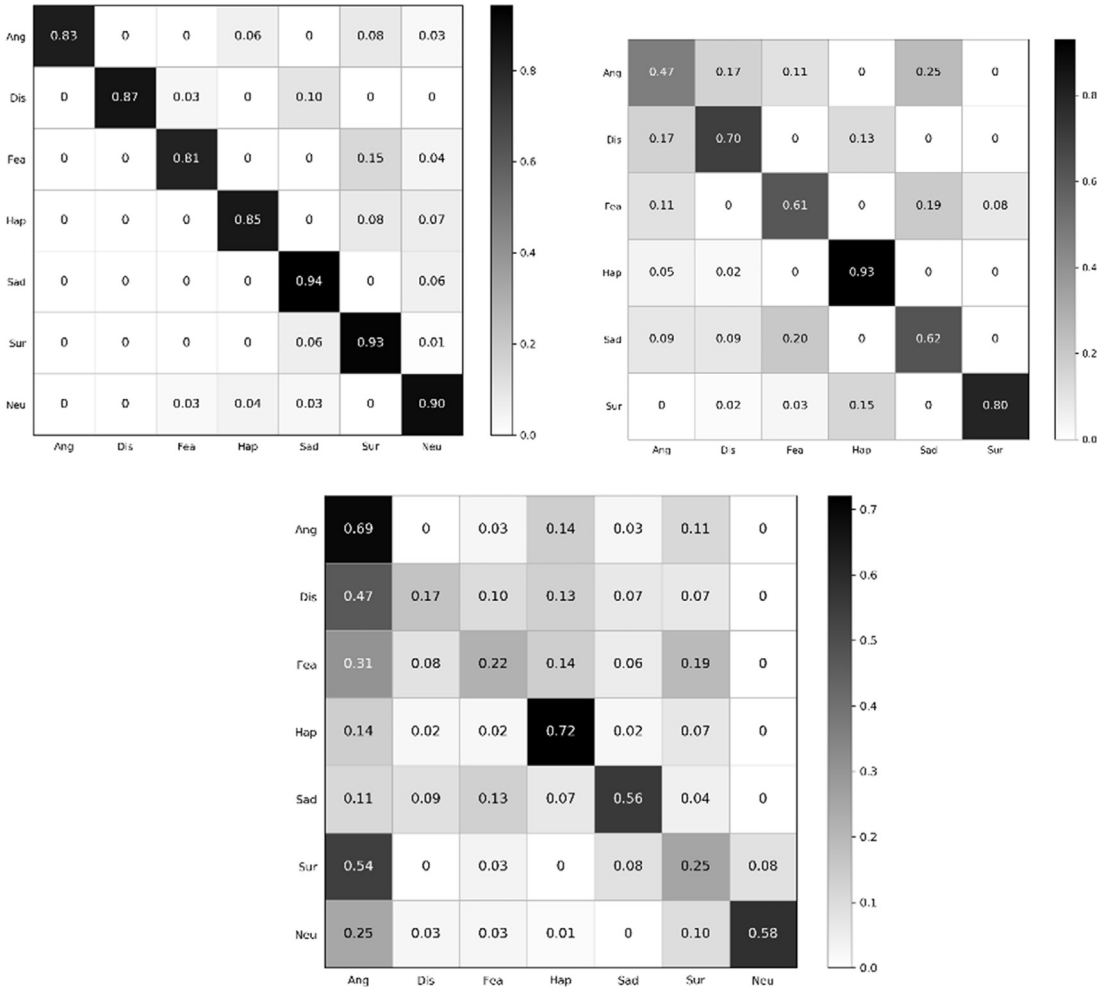


Fig. 7. Confusion matrices of VGG-CTSLSTM on CK+, MMI and AFEW (from left to right).

Table 5
The relationship between number of segments and accuracy.

Segments	Frames	Accuracy (MMI, CK+)
4	5	62.4, 81.3
5	4	66.5, 86.4
7	3	**68.8**, **87.6**

The first experiment aims to investigate the effectiveness of the LEMHI features, it is conducted on the CK+ dataset. As the performance of LEMHI might rely on the size of landmarks window and the value of enhancement factor α, we conducted our experiment using different sizes and α. As is shown in Table 4, coefficient 2 has a better performance than other coefficients. Also, using windows with size 20×20 achieves the best performances 83.22.

The second experiment compares the performance difference between the MHI and LEMHI features. The recognition rates in using MHI and LEMHI which employs facial landmarks are summarized in Fig. 6. As is shown, no matter MMI or CK+, the performance of LEMHI-CNN model in each database obviously better than that of the MHI-CNN model.

The third experiment investigates the effectiveness of our CNN-RNN model on three different databases, Fig. 7 shows the resulting confusion matrices of our model on these databases. We also explore different LSTM parameter settings, including the different number of LSTM net's layers, different number of LSTM's hidden units. Due to the limited training data, we find that less layers and hidden units achieve better accuracy, which means simpler structure has better classification ability. Since the number of segments when we apply cross temporal segment to CNN features will influence the final accuracy, an experiment is carried out to explore the relationships between them. Results are shown in Table 5, frames means the number of frame in one segments, when there are 5 frames in each segment, all 16 frames can be divided into 4 segments due to one frame overlapping. It is obvious that less segments with more frames in one segment result in higher accuracy probably because it preserves more spatial features.

In the last experiment we test our integrated framework on three datasets. Table 6 shows the results using two individual models separately, and the final accuracy achieved by integrated framework. Proposed framework achieves 93.9%, 78.4%, 51.2% accuracy on CK+, MMI, AFEW respectively. Comparing to other state-of-the-art works, our framework outperforms others in CK+ and AFEW databases while achieves comparable results in MMI.

According to [37] and [13], all of the final results are based on audio-video emotion recognition, that is to say, audios of AFEW datasets are used to train an audio modality in [37,13], which has been proved to bring ~3% gain in recognition accuracy. In [37], the author trained an SVM with the RBF kernel using audio features extracted with the OpenSmile toolkit. Their method achieve performance 49% using linear SVMs, without combining with audio features while achieve 53.8% recognition accuracy after incorporating audio features. When it comes to [13], an accuracy of 48.30% can be reached if the two models (CNN-RNN and C3D)

Table 6
Accuracy of LEMHI-VGG, VGG-CTSLSTM and Integrated framework respectively evaluated on CK+, MMI and MMI.

Datasets	State of art method	VGG-CTSLSTM	LEMHI-VGG	Fusion
CK+	84.1 [30], 84.4 [31], 88.5 [32], 92.4 [33], 93.2 [19]	87.6	83.2	**93.9**
MMI	63.4 [33], 75.12 [34], 86.7 [35], 78.51 [36], 77.5 [19]	68.8	66.5	78.4
AFEW	53.8 [37], 37.6 [38], 59.02 [13]	45.6	43.7	**51.2**

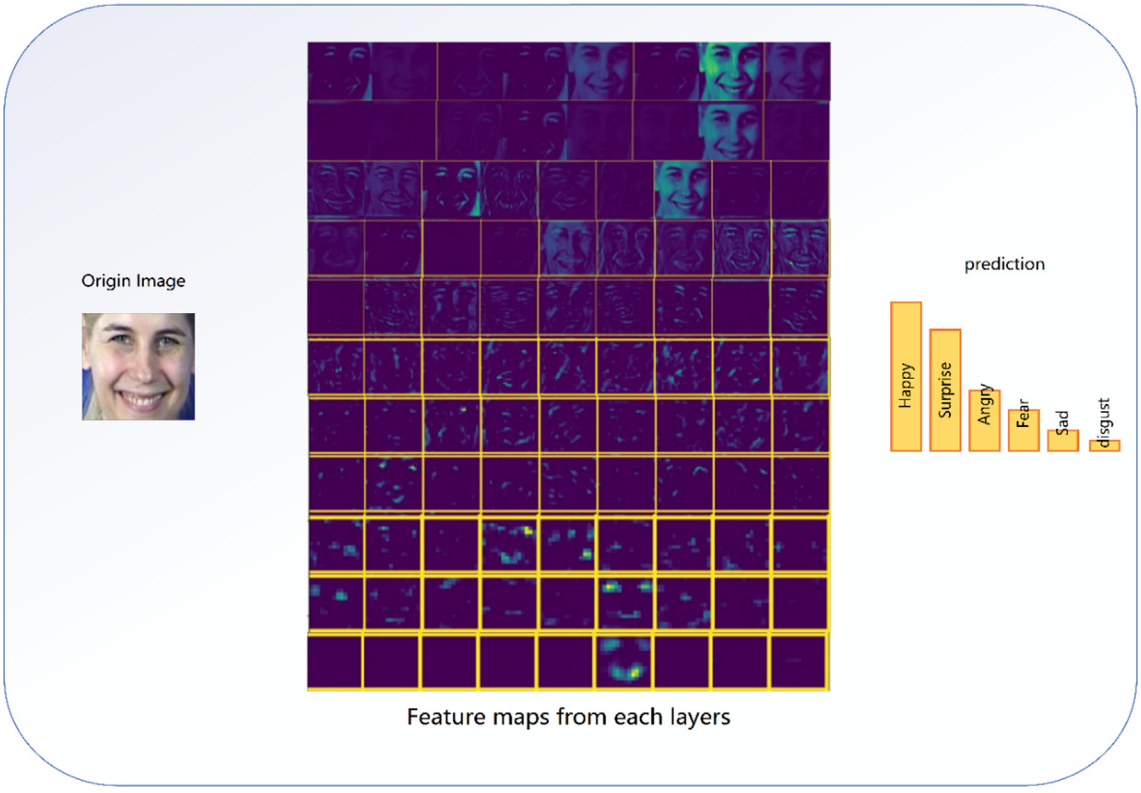


Fig. 8. Visualization of feature maps from VGG CovNet.

are fused without audio information. However, the final recognition accuracy rises to 59.02% due to extra audio information. In contrast, we conducted our experiment without using audios. Only frames come from AFEW video are extracted to train our framework.

As for [35], their boosted LBP based SVM approach achieve impressive accuracy of 86.7%, which is much higher than our proposed method, this situation may result from a different recognition task in [35], where one neural face and three peak frames of each sequence are picked manually as input of static image emotion recognition. As a contrast, several frames randomly captured from emotion sequences as input to our model. Thus, we can also conclude that the combination of two models improves the recognition rate.

Besides recognition accuracies, we would like to attain further insight into the learned CovNet models. In this sense, we visualize every layer activation in order to decipher which parts of the face yield the most discriminative information. As is shown in Fig. 8, the initial volume stores the raw image pixels and the last volume stores the class scores. In first few convolutional layers, the profile of human face has been extracted. Although these features become more abstract when it comes to deeper layers, it shows that activated areas in feature maps have strong correlations to facial AUs, which is also known as crucial portions of face, and it may explain why our local enhanced MHI, which puts intensities on these facial AUs aids in improving the performance of deep networks.

5. Conclusion

This paper presents a facial expression recognition framework using LEMHI-CNN and CNN-RNN. The integrated framework incorporates facial landmarks to enable attention-aware facial motion capturing and utilize neural networks to extract spatial-temporal features and classify them, which achieves better performance than most of the state-of-the-art methods on CK+, MMI and AFEW dataset. Our main contributions are threefold. Firstly, we proposed an attention-aware facial motion features based on MHI. Secondly, we introduced temporal segment LSTM to video emotion recognition and improve it. Thirdly, we integrated two models with late fusion based on random weight search.

Although the accuracy we achieved on CK+ and MMI dataset is comparably good, the accuracy on AFEW dataset is still unsatisfactory. That is to say, current methods can be employed in real world. However, how to improve the performance on wild expression dataset, such as AFEW, needs to be further explored.

Acknowledgement

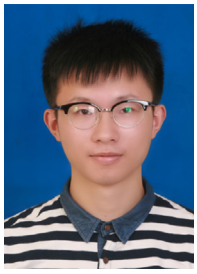
This research has been partially supported by National Natural Science Foundation of China (Grant No. 61672202, 61502141), State Key Program of NSFC-Shenzhen Joint Foundation (Grant No. U1613217) and State Key Program of National Natural Science of China (61432004).

References

[1] R. Ranjan, S. Sankaranarayanan, A. Bansal, et al., Deep learning for understanding faces: machines may be just as good, or better, than humans, IEEE Signal Process Mag. 35 (1) (2018) 66–83.
[2] W. Liu, Y. Wen, Z. Yu, et al., Sphere face: deep hypersphere embedding for face recognition, arXiv preprint arXiv:1704.08063, 2017.
[3] C. Liu, H. Wechsler, Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition, IEEE Trans. Image Process. 11 (4) (2002) 467–476.
[4] T. Ahonen, A. Hadid, M. Pietikainen, Face description with local binary patterns: Application to face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 12 (2006) 2037–2041.
[5] X. Qi, R. Xiao, C. Li, Y. Qiao, J. Guo, X. Tang, Pairwise rotation invariant co-occurrence local binary pattern, TPAMI 36 (11) (2014) 2199–2213.
[6] S. Leutenegger, M. Chli, R. Siegwart, BRISK: Binary robust invariant scalable keypoints. In ICCV, pages 2548–2555, 2011.
[7] D. Yueqi, L. Jiwen, W. Ziwei, et al., Learning Deep Binary Descriptor with Multi-Quantization[J], IEEE Trans. Pattern Anal. Mach. Intell. (2018), 1–1.
[8] J. Lu, V.E. Liong, X. Zhou, et al., Learning compact binary face descriptor for face recognition, IEEE Trans. Pattern Anal. Mach. Intell 37 (10) (2015) 2041–2056.
[9] Y. Duan, J. Lu, J. Feng, et al., Context-aware local binary feature learning for face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 40 (5) (2018) 1139–1153.
[10] J. Lu, V.E. Liong, J. Zhou, Simultaneous local binary feature learning and encoding for homogeneous and heterogeneous face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 40 (8) (2018) 1979–1993.
[11] M. Wang, W. Deng, Deep face recognition: a survey, arXiv preprint arXiv:1804.06655, 2018.
[12] Y. LeCun, F.J. Huang, L. Bottou, Learning methods for generic object recognition with invariance to pose and lighting, in: Computer Vision and Pattern Recognition, 2004, CVPR 2004, Proceedings of the 2004 IEEE Computer Society Conference on. IEEE, 2004, 2: II-104.
[13] Y. Liu et al., Video-based emotion recognition using CNN-RNN and C3D hybrid networks, in: ACM International Conference on Multimodal Interaction, ACM, 2016, pp. 445–450.
[14] J. Donahue, L.A. Hendricks, M. Rohrbach, et al., Long-term recurrent convolutional networks for visual recognition and description, IEEE Trans. Pattern Anal. Mach. Intell. 39 (4) (2017) 677–691.
[15] B.K. Kim, H. Lee, J. Roh, et al., Hierarchical committee of deep CNNs with exponentially-weighted decision fusion for static facial expression recognition, in: ACM on International Conference on Multimodal Interaction, ACM, 2015, pp. 427–434.
[16] G. Pons, D. Masip, Multi-task, multi-label and multi-domain learning with residual convolutional networks for emotion recognition, arXiv preprint arXiv:1802.06664, 2018.
[17] S. Hosseini, S.H. Lee, N.I. Cho, Feeding hand-crafted features for enhancing the performance of convolutional neural networks, arXiv preprint arXiv:1801.07848, 2018.
[18] S. Koelstra, M. Pantic, I. Patras, A dynamic texture-based approach to recognition of facial actions and their temporal models, IEEE Trans. Pattern Anal. Mach. Intell. 32 (11) (2010) 1940–1954.
[19] B. Hasani, M.H. Mahoor, Facial expression recognition using enhanced deep 3D convolutional neural networks, in: Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on. IEEE, 2017, pp. 2278–2288.
[20] M. Schuster, K.K. Paliwal, Bidirectional recurrent neural networks, IEEE Trans. Signal Process. 45 (11) (2002) 2673–2681.
[21] Y.G. Kim, X.P. Huynh, Discrimination between genuine versus fake emotion using long-short term memory with parametric bias and facial landmarks, in: IEEE International Conference on Computer Vision Workshop, IEEE Computer Society, 2017, pp. 3065–3072.
[22] C.Y. Ma, M.H. Chen, Z. Kira, et al., TS-LSTM and temporal-inception: exploiting spatiotemporal dynamics for activity recognition, 2017.
[23] J. Bergstra, Y. Bengio, Random search for hyper-parameter optimization, J. Mach. Learning Res. 13 (1) (2012), pp. 281–305. 5.
[24] J.W. Davis, Appearance-based motion recognition of human actions, Massachusetts Institute of Technology (1996).
[25] A. Dhall, R. Goecke, S. Lucey, et al., Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark, in: IEEE International Conference on Computer Vision Workshops, IEEE, 2011, pp. 2106–2112.
[26] I.J. Goodfellow, D. Erhan, P.L. Carrier, et al., Challenges in representation learning: A report on three machine learning contests, Neural Networks 64 (2015) 59–63.
[27] M. Pantic, M. Valstar, R. Rademaker, et al., Web-based database for facial expression analysis, in: 2005 IEEE international conference on multimedia and Expo, IEEE, 2005, p. 5.
[28] P. Lucey, J.F. Cohn, T. Kanade, et al., The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression, in: Computer Vision and Pattern Recognition Workshops, IEEE, 2010, pp. 94–101.
[29] A. Dhall, R. Goecke, S. Lucey, et al., Collecting large, richly annotated facial-expression databases from movies, IEEE Multimedia 19 (3) (2012) 34–41.
[30] C. Mayer, M. Eggers, B. Radig, Cross-database evaluation for facial expression recognition, Pattern Recognit Image Anal. 24 (1) (2014) 124–132.
[31] S.H. Lee, M.R. Yong, Intra-class variation reduction using training expression images for sparse representation based facial expression recognition, IEEE Trans. Affective Comput. 5 (3) (2017) 340–351.
[32] S. Taheri, Q. Qiu, R. Chellappa, Structure-preserving sparse decomposition for facial expression analysis, IEEE Trans Image Process 23 (8) (2014) 3590–3603.
[33] M. Liu, S. Li, S. Shan, et al., Deeply learning deformable facial action parts model for dynamic expression analysis, in: Asian Conference on Computer Vision, Springer, Cham, 2014, pp. 143–157.
[34] M. Liu, S. Shan, R. Wang, et al., Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, 2014, pp. 1749–1756.
[35] C. Shan, S. Gong, P.W. Mcowan, Facial expression recognition based on Local Binary Patterns: A comprehensive study, Image Vis. Comput. 27 (6) (2009) 803–816.
[36] M.R. Mohammadi, E. Fatemizadeh, M.H. Mahoor, PCA-based dictionary building for accurate facial expression recognition via sparse representation, J. Vis. Commun. Image Represent 25 (5) (2014) 1082–1092.

[37] X. Fan, T. Tjahjadi, A dynamic framework based on local Zernike moment and motion history image for facial expression recognition, Pattern Recogn. 64 (2017) 399–406.
[38] A. Yao, J. Shao, N. Ma, et al., Capturing AU-aware facial features and their latent relations for emotion recognition in the wild, in: ACM on International Conference on Multimodal Interaction, ACM, 2015, pp. 451–458.
[39] J. Schmidhuber, Multi-column deep neural networks for image classification, in: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, 2012, pp. 3642–3649.

Min Hu received the M.S. degree in industrial automation from Anhui University, China, in 1994, and the Ph. D.degree in computer science from Hefei University of Technology, Hefei, China, in 2004. She is currently a Professor with School of Computer and Information, Hefei University of Technology. His research interests include digital image processing, artificial intelligence and data mining.



Haowen Wang received B.S. degree from Anhui University, Hefei, China, in 2016. He is currently a master student in Hefei University of Technology, Hefei, China. His research interests include computer vision and machine learning.

Xiaohua Wang received the Ph.D. degree in computer science from Hefei Institute of Physical Science, Chinese Academy of Sciences, China, in 2005. She is currently an associate professor with School of Computer and Information, Hefei University of Technology. Her research interests include affective computing, artificial intelligence and visual pattern recognition.

Ronggui Wang received the M.S. degree in mathematics from Anhui University, China, in 1997, and the Ph.D. degree in computer science from Hefei University of Technology, Hefei, China, in 2005. He is currently a Professor with School of Computer and Information, Hefei University of Technology. His research interests include digital image processing, artificial intelligence and data mining.

Juan Yang received the B.S. and M.S. degrees in mathematics from Hefei University of Technology, Hefei, China, in 2004 and 2008, respectively. She received the Ph.D. degree with school of Computer and Information, Hefei University of Technology. She is currently a lecturer with school of Computer and Information, Hefei University of Technology. Her research interests include image processing and intelligent visual surveillance.