doi:10.3772/j.issn.1006-6748.2023.02.003

Video expression recognition based on frame-level attention mechanism①

CHEN Rui(陈 瑞)*, TONG Ying②*, ZHANG Yiye**, XU Bo**
(*College of Information & Communication Engineering, Nanjing Institute of Technology, Nanjing 211167, P. R. China)
(**Jiangsu Future Network Innovation Research Institute, Nanjing 211111, P. R. China)

Abstract

Facial expression recognition (FER) in video has attracted the increasing interest and many approaches have been made. The crucial problem of classifying a given video sequence into several basic emotions is how to fuse facial features of individual frames. In this paper, a frame-level attention module is integrated into an improved VGG-based frame work and a lightweight facial expression recognition method is proposed. The proposed network takes a sub video cut from an experimental video sequence as its input and generates a fixed-dimension representation. The VGG-based network with an enhanced branch embeds face images into feature vectors. The frame-level attention module learns weights which are used to adaptively aggregate the feature vectors to form a single discriminative video representation. Finally, a regression module outputs the classification results. The experimental results on CK + and AFEW databases show that the recognition rates of the proposed method can achieve the state-of-the-art performance.

Key words: facial expression recognition (FER), video sequence, attention mechanism, feature extraction, enhanced feature, VGG network, image classification, neural network

0 Introduction

Facial expression is one of the most powerful and natural signals for human beings to convey their emotional states and intentions. Numerous researches have been conducted on automatic facial expression recognition (FER) because it's practical importance in fatigue driving, online teaching, telemedicine and many other artificial intelligent systems. Video-based FER systems aim at classifying expressions into seven emotions, i.e. happy, angry, disgust, fear, sad, neutral, and surprise. The popular video-based FER system mainly includes three modules, namely frame pre-processing, feature extraction, and classification[1]. Specifically, frame pre-processing module refers to face detection, alignment, illumination normalizing, etc. Feature extraction and classification are two key modules, in which feature extraction module is crucial for improving the recognition accuracy. Feature extraction module encodes video frames into compact feature vectors, then these vectors are subsequently fed into the classification module for prediction.

Feature extraction methods for video-based FER can be roughly divided into two types: static-based methods and spatial-temporal methods. In static-based methods[2-3], the feature representation is encoded with only spatial information from a single image. These methods use handcrafted or learned features. Optical flow method uses the changes of pixels in time domain and the correlation between sequence frames to mine the corresponding relationship between adjacent frames[4]. It is widely used to extract facial expression and motion information in video sequences. Hidden Markov model[5] abstracts expression features into state sequences, and uses probability matrix to describe the emergence and transition of states, so as to realize expression training and recognition. As a vision based template method, motion history image (MHI)[6] is applied to FER. This method shows the target motion in the form of image brightness by calculating the pixel changes at the same position in a certain time period. In recent years, many researchers have also tried to extend the feature extraction method with better recognition effect in static images to dynamic features, such as local binary pattern (LBP) on three orthogonal planes (LBP-TOP)[7]. LBP-TOP extracted LBP features from the Gabor filtered image, then extended the features to three dimensions by adding time dimension. Discriminant graph regularized non-negative matrix factorization

① Supported by the Future Network Scientific Research Fund Project of Jiangsu Province (No. FNSRFP2021YB26), the Jiangsu Key R&D Fund on Social Development (No. BE2022789), and the Science Foundation of Nanjing Institute of Technology (No. ZKJ202003).
② To whom correspondence should be addressed. E-mail: tongying@njit.edu.cn.
Received on Sep. 9, 2022

(DGNMF)[8] encoded the geometrical class information by constructing an affinity graph to obtain a better representation. Ref. [9] proposed an enhanced dictionary pair learning sparse representation (EDPLSR) framework which jointly learned a synthesis dictionary as well as an analysis dictionary. They introduced a manifold regularization term to obtain a smooth and sparse representation along the geodesics of data manifold.

Since 2013, emotion recognition competitions, such as FER 2013[10] and Emotion Recognition in the Wild (EmotiW)[11] have collected relatively sufficient training data from challenging real-world scenarios, which implicitly promotes the transition of FER from lab-controlled to in-the-wild settings. In the meanwhile, due to the dramatically increased chip processing abilities (e. g., graphics processing unit(GPU)) and well-designed network architecture, studies in various fields have begun to transfer to deep learning methods, which have achieved the state-of-the-art recognition accuracy and exceeded previous results by a large margin. Deep learning attempts to capture high-level abstractions through hierarchical architectures of multiple nonlinear transformations and representations. Convolutional neural network (CNN) has been extensively used in diverse computer vision applications, including FER. CNN is robust to face location changes and scale variations and behaves better than the multilayer perceptron (MLP) in the case of previously unseen face pose variations. For example, Li and Zhang[12] took advantage of deep CNN for feature extraction and won the FER 2013. Jung et al.[13] proposed a joint fine-tuning network method based on two different models to improve the recognition accuracy. Mundher et al.[14] used deep CNN to learn facial expression features. Likewise, given with more effective training data of facial expression, deep learning techniques have been increasingly implemented to handle the challenging factors for emotion recognition in the wild.

Aiming at modelling the temporal or motion features in video sequences, spatial-temporal methods are presented in Refs [15-17]. Two widely used spatial-temporal methods for video-based FER are long short-term memory (LSTM)[16] and C3D[18]. LSTM derives temporal information from video sequences by exploiting the fact that feature vectors are connected semantically for contiguous frames. So, various CNN-LSTM models in Refs [19, 20] are proposed to obtain spatial-temporal information to improve recognition accuracy. In view of the limited representation ability of single-layer LSTM and the limitation of its generalization ability when solving complex problems, Irsoy and Cardie[21] pointed out that LSTM with multi-layer architecture can achieve better results. Furthermore, Sutskever et al.[22] developed a 4-layer LSTM to achieve good machine translation performance.

Among all the above methods, static-based methods can achieve better recognition rate according to several winner solutions in EmotiW challenges. For video-based FER, because the frames in a given video clip may vary in expression intensity, directly measuring per-frame error does not yield satisfactory performance. Various spatial-temporal methods need a frame aggregation operation to obtain a video-level result. The most convenient way is to directly concatenate the output of these frames, for example, Samira et al.[23] concatenated the n-class probability vectors of 10 segments to form a fixed-length video representation by frame averaging. However, the number of frames in each sequence may be different. Two approaches have been considered to generate a fixed-length feature vector for each sequence in Refs [24,25]. Sarah et al.[26] proposed a statistical encoding module to aggregate frame features which computes the mean, variance, minimum, and maximum of the frame feature vectors. These methods have the limitation of ignoring the importance of frames for FER. Since the attention mechanism can quickly focus on regions of interest in complex scenes, Xu et al.[27] introduced attention mechanism in the image annotation, and calculated the weight for each region of the input sequence at different times of decoding, and then focused on different image regions to generate more reasonable words. Liu et al.[28] used self-attention (SA) mechanism to update learning parameters only through its own information.

Inspired by SA mechanism, a frame-level attention mechanism based video expression recognition method is proposed, abbreviating as ECNN-FSA. First, a frame-based self-attention mechanism is proposed to learn SA kernels and relation attention kernels for frame importance reasoning. For feature extraction, the fully connected layer is removed from the traditional VGG-16 and an enhanced branch is introduced from the seventh layer to extract more abundant facial expression features. The contributions of this paper are as follows.

(1) A clear structure of enhanced CNN with frame-level self-attention mechanism is proposed to solve video-based FER problem, denoted by ECNN-FSA.

(2) The traditional VGG-16 network with an enhanced branch is improved to obtain hierarchical facial expression features and enrich expression information.

(3) A frame-level self-attention mechanism for adaptively frame feature aggregating is proposed to form a single discriminative video expression.

The rest of this paper is organized as follows. In Section 1, the proposed ECNN-FSA framework is introduced, including model architecture, the improved VGG-16 network, and frame-level attention module. In Section 2, the experimental results on CK + and AFEW databases are discussed. Section 3 is the conclusion of this paper.

1 Proposed ECNN-FSA framework

1.1 Model architecture

The framework of the proposed ECNN-FSA video expression method consists of three modules: feature extraction module, FSA module and regression module, which is depicted in Fig. 1. It takes a facial video with a variable number of face images as its input and produces a fixed-dimension feature representation for FER. The feature extraction module is a deep hierarchical spatial feature extractor based on VGG-16 with an enhanced feature branch. Different from Ref. [16], the fully connected (FC) layer is replaced with global average pooling (GAP) layer. The extracted features are input into FSA module to get the attention weight W_{sa} by calculating the correlation between feature vectors. The features are fused by W_{sa} and then input into the regression module. The first layer of the regression module is the FC layer for feature mapping. ReLU activation layer is used to ensure the nonlinearity of FC layer. The dropout layer randomly 'inactivates' the eigenvalues to avoid over-fitting. The FC layer is used to feature mapping and the Softmax layer outputs the probability of each expression.

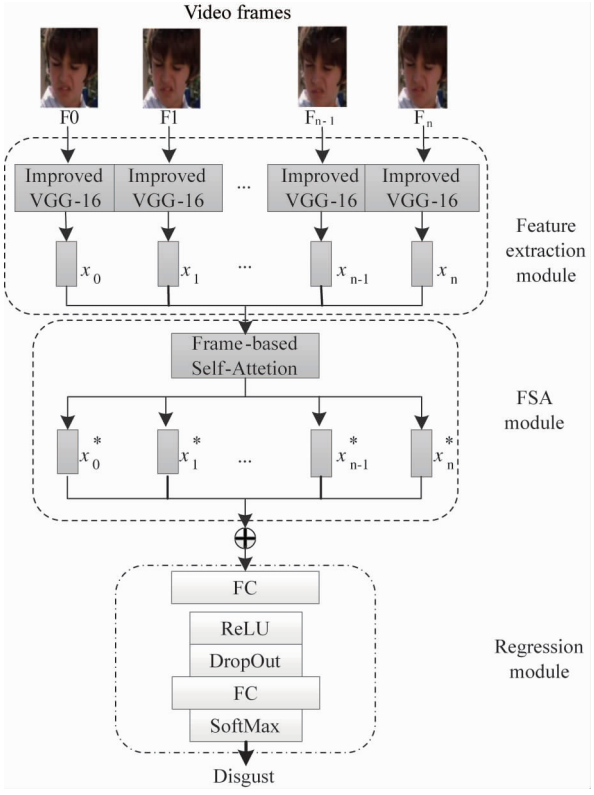

Fig. 1 Proposed ECNN-FSA framework

Formally, a video $\boldsymbol{V}$ with m frames is denoted as $\boldsymbol{V} = [F_1, F_2, \cdots, F_m]$, and the facial frame features is denoted as $\boldsymbol{X} = [\boldsymbol{x}_0, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_m]$. In the specific implementation, as shown in Fig. 1, the n ($n \leqslant m$) consecutive video frames are processed at a time, and the corresponding feature vectors are got through the feature extraction module. The FSA module outputs distinctive features $[\boldsymbol{x}_0^*, \boldsymbol{x}_1^*, \cdots, \boldsymbol{x}_n^*]$ that are fused into the regression module. The first FC layer is used to map the learned deep semantic features into the sample label space for classification.

1.2 Improved VGG-16 network

The feature extraction module is based on VGG-16 network. As shown in Fig. 2, the fully connected (FC) layers are replaced with a global average pooling (GAP) layer to reduce complexity, and three convolutional layers (see the dotted box in Fig. 2) are added to obtain deeper expression semantic information. The kernels of the three added convolutional layers are $3 \times 3 \times 512 \times 1024$. Then, an enhanced branch is introduced to obtain larger receptive field, which consists of a single convolutional layer (kernel size is $7 \times 7 \times 256 \times 1024$) and a GAP layer. The GAP layer can not only reduce the calculation of model parameters, but also make the output feature dimension just related to the number of channels of feature mapping, and has nothing to do with the size of feature mapping. Therefore, the authors can process images of any size and the output feature dimensions are kept consistent.

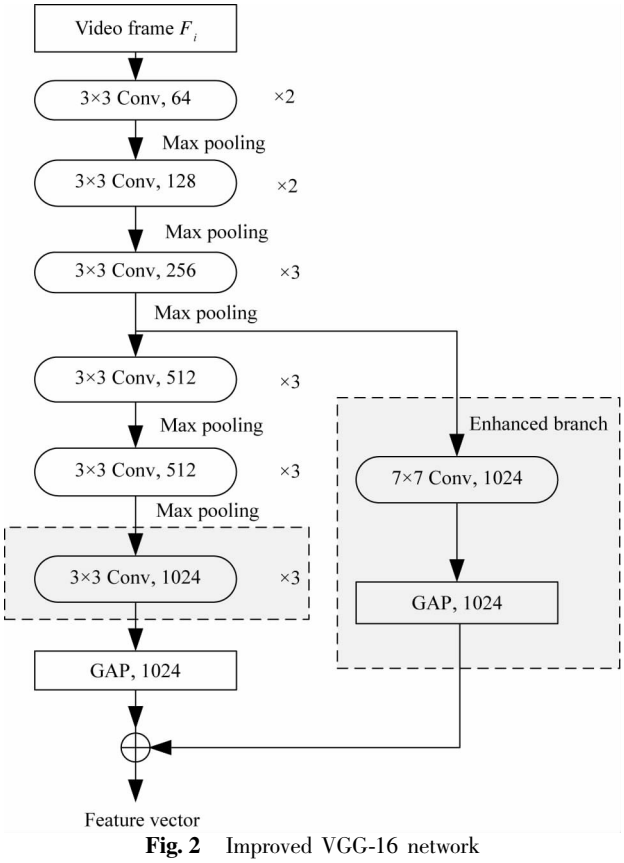

Fig. 2 Improved VGG-16 network

1.3 Frame-level self-attention module

Human visual system uses attention mechanism to screen information, which can quickly locate regions of interest in complex scenes. Recently, Ref. [29] proposed SA mechanism, which achieved excellent results in natural language processing. Fajtl et al.[30] introduced SA mechanism into video processing. They scored each video frame according to the correlation between frames, so as to obtain the key frames of video. Inspired by these researches, SA module is used to learn the internal dependence between video frames, and capture the internal structure, and so obtain the significant characteristics of differentiation. On the other hand, SA mechanism is a kind of mean operation, which can effectively avoid the gradient loss problem caused by the deepening of network layers and greatly speed up the network training speed.

Attention mechanism can be described as mapping a query and a set of key-value pairs to an output. As described in Ref. [29], the attention function is calculated on a set of queries simultaneously and packed together into a matrix $\boldsymbol{Q}$. The keys and values are packed together into matrices $\boldsymbol{K}$ and $\boldsymbol{V}$. The outputs matrix of scaled dot-product attention are computed as

$$Attention(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{1}$$

where, d_k is the dimension of $\boldsymbol{Q}$ or $\boldsymbol{K}$, and $\boldsymbol{Q} \in R^{n \times d_k}$, $\boldsymbol{K} \in R^{m \times d_k}$, $\boldsymbol{V} \in R^{m \times d_v}$. $\sqrt{d_k}$ is used to scale the dot products. SA can be regarded as a special case of attention mechanism where $\boldsymbol{K} = \boldsymbol{V} = \boldsymbol{Q}$. It can capture the internal structure of the sequence by learning the dependency of the sequence content with less computation complex. In this paper, FSA mechanism is used to focus on the frame with the greatest difference in video sequences and distinguish the frame most related to the video expression classification, which is shown in Fig. 3.

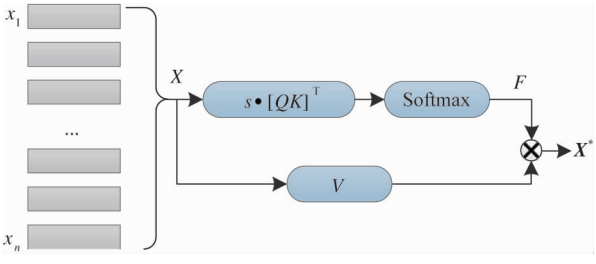

Fig. 3 FSA module

In Fig. 3, $\boldsymbol{X} = [x_0, x_1, \cdots, x_n]$ is the feature vector of n consecutive frames output from the improved VGG-16. Matrices $\boldsymbol{Q}$, $\boldsymbol{K}$ and $\boldsymbol{V}$ are calculated by

$$\begin{cases} \boldsymbol{Q} = \boldsymbol{W}_q \boldsymbol{X} \\ \boldsymbol{K} = \boldsymbol{W}_k \boldsymbol{X} \\ \boldsymbol{V} = \boldsymbol{W}_v \boldsymbol{X} \end{cases} \tag{2}$$

where, $\boldsymbol{W}_q$, $\boldsymbol{W}_k$ and $\boldsymbol{W}_v$ are weight matrices of different network. The attention weight matrix $\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}$ describes the correlation between elements in the input feature matrix $\boldsymbol{X}$. The super parameter s is manually set to suppress the attention weight value, which is set to 0.1 in this paper. The Softmax function is used to normalize the attention weight, and then multiply it with $\boldsymbol{V}$ to get the differentiated saliency matrix $\boldsymbol{X}^*$. The weight matrices of $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ are 2048×2048, and the computation complex is lower when compared with the fully connected layer.

2 Experiments

2.1 Databases and implementation details

CK +: the Extended Cohn-Kanade (CK +) database[31] is the most extensively used laboratory-controlled database for evaluating FER systems. CK + database contains 593 video sequences from 123 subjects. The sequences vary in duration from 10 to 60 frames and show a shift from a neutral facial expression to the peak expression. Among these videos, 327 sequences from 118 subjects are labeled with seven basic expression labels (i.e. anger, contempt, disgust, fear, happiness, sadness, and surprise) based on the Facial Action Coding System (FACS). Because CK + does not provide specified training, validation and test sets, the algorithms evaluated on this database are not uniform. For static-based methods, the most common data selection method is to extract the last one to three frames with peak formation and the first frame (neutral face) of each sequence. Then, the subjects are divided into n groups for person-independent m-fold cross-validation experiments, where commonly value of m can be selected in [5, 8, 10]. Some examples of aligned images under laboratory environments in CK + are shown in Fig. 4.

Fig. 4 Examples of aligned images under laboratory environments (CK +)

AFEW: the Acted Facial Expressions in the Wild (AFEW) database was established and introduced in Ref. [32] and has served as an evaluation platform for the annual challenge EmotiW since 2013. AFEW database contains video clips collected from different movies with spontaneous expressions, various head poses, occlusions and illuminations. It is a temporal and multimodal database that provides with vastly different environmental conditions in both audio and video. Samples are labeled with seven expressions: anger, disgust, fear, happiness, sadness, surprise, and neutral. The annotation of expressions have been continuously updated, and reality TV show data have been continuously added. The AFEW 7.0 in EmotiW 2017[33] is divided into three data partitions in an independent manner in terms of subject and movie/TV source: Train (773 samples), Val (383 samples) and Test (653 samples), which ensures data in the three sets belong to mutually exclusive movies and actors. Some examples of images under real-world conditions in AFEW are shown in Fig. 5.

Fig. 5 Examples of images under real-world conditions (AFEW)

Implementation details: the hardware configurations of the experiments are: an Intel Core i7-7800X CPU, 2 NVIDA GeForce GTX 1080Ti GPUs. This work implements the proposed method by the Pytorch toolbox. The video frames are preprocessed by face detection and alignment in the Dlib toolbox, and the face bounding box is extended with a ratio of 25% and then the cropped faces are resized to scale of 224×224. For feature extraction, the pre-trained model weights on SFEW and FER2013 databases are loaded into the improved VGG-16. The initial learning rate of the model is set to 0.001, which decreases during the training process.

For training, on both CK + and AFEW, the authors set a batch to have 48 instances with n frames in each instance. For frame sampling in a video, first split on frame from each segment. The default value of n is set to 5. The stochastic gradient descent (SGD) algorithm is used for optimization with a momentum of 0.9. The parameters used in the proposed ECNN-FSA model are shown in Table 1.

To eliminate the influence of complex background on facial expression recognition, the video images in AFEW and CK + databases are pre-processed similar to Ref. [16]. The proposed network is pre-trained and fine-tuned on AFEW database. VGG-FACE weights are used as the initial weight of backbone CNN. Then, partial samples from SFEW and FER 2013 are used to fine tune the proposed ECNN-FSA. The best network parameters are obtained by using the training set of AFEW and the expanded training samples.

Table 1 Specific parameters setting

Module	Output	Parameter
Self-attention	$n \times 2048$	$\boldsymbol{K}(2048 \times 2048)$, $\boldsymbol{V}(2048 \times 2048)$, $\boldsymbol{Q}(2048 \times 2048)$, $s(0.1)$
Layer-norm	2048	-
ReLU	2048	-
Dropout	2048	0.5
FC	7	-
Softmax	7	-

2.2 Ablation study

Discussion on cascade LSTM. To explore the role of the number of cascade layers of LSTM, this work has performed an ablation study on multi-layer prediction. It can be seen from Table 2 that the recognition performance of CNN with 2-layer LSTM outperforms CNN with only single-layer LSTM in terms of F1 score and accuracy. The values in parentheses represent the dimensionality of the feature vectors output by each LSTM.

Table 2 Experimental results of LSTM networks with different number of layers and parameters on AFEW database

Model	F1 score	Accuracy
CNN-LSTM (2048)	0.2895	33.69%
CNN-LSTM (3000)	0.2954	32.88%
CNN with 2 layer LSTMs (3000, 3000)	0.3069	**34.77%**
CNN with 2 layer LSTMs (2048, 2048)	**0.3279**	34.50%
CNN with 2 layer LSTMs (2048, 1024)	0.2950	34.23%

Discussion on feature extraction. To illustrate the advantages of the enhanced CNN, this work conducts comparative experiments with CNN. As shown in Table 3, the performance of ECNN with 2 layer LSTMs (Fc1, 7×7) is the best, which shows that the proposed ECNN can improve the feature extraction performance. The parameter 'Fc1' in parentheses indicates one fully connected layer aggregated in the enhanced branch, and '5×5' represents the kernel size of the first convolution layer of the enhanced branch.

Table 3 Experimental results of ECNN-LSTM on AFEW database

Model	F1 score	Accuracy
Baseline	-	38.81%
ECNN with 2 layer LSTMs (Fc1, 5×5)	0.3733	40.16%
ECNN with 2 layer LSTMs (Fc1, 7×7)	0.3816	41.25%
ECNN with 2 layer LSTMs (Fc2, 5×5)	0.3514	39.34%
ECNN with 2 layer LSTMs (Fc2, 7×7)	0.3763	40.44%

2.3 Evaluation on CK + database

The proposed ECNN-FSA method is evaluated on CK + with comparison to several state-of-the-art methods in Table 4, i.e. 3DCNN-DAP[4], LOMo[34], STM-ExpLet[28], DTAGN[13], and CNN-LSTM[30]. The testing results are obtained after 5 times of cross validation. Due to the fact that the videos in CK + database show a shift from a neural facial expression to the peak expression, these methods conduct data selection manually. LOMo[34] uses all frames with a new latent ordinal model which extracts CNN features for sub-event detection and uses multi-instance SVM to classify the facial expression. STM-ExpLet[28] combines a spatial CNN model and a temporal network, where the spatial CNN model only uses the last peak frame. DTAGN[13] selects a fixed length sequence for each video with a lipreading method. CNN-LSTM[30] uses the last three frames and the first frame for each video. The baseline method uses the improved VGG-16 to generate scores for individual frame and applied score fusion for all frames. It achieves 94.6% which is better than LOMo[34]. The proposed ECNN-FSA achieves the highest accuracy rate 97.95%, which is 5.6%, 4.07%, 1.52% and 2.03% higher than 3DCNN-DAP, STM-ExpLet, DTAGN, and CNN-LSTM, respectively. Table 5 shows the corresponding confusion matrix.

Table 4 Evaluation of ECNN-FSA with a comparison to state-of-the-art methods on CK + database

Method	Accuracy/%
3DCNN-DAP[4]	92.35
LOMo[34]	92.00
STM-ExpLet[28]	93.88
DTAGN[13]	96.43
CNN-LSTM[30]	95.92
Score fusion (baseline)	94.6
Proposed ECNN-FSA	97.95

Table 5 Confusion matrix of ECNN-FSA network on CK + database (%)

	An	Di	Fe	Ha	Ne	Sa	Su
An	**100**	0	0	0	0	0	0
Di	0	**94.16**	0	4.41	0	0	1.44
Fe	0	0	**98.3**	0	1.7	0	0
Ha	0	1.44	0	**97.12**	0	1.44	0
Ne	0	0	0	0	**98.5**	0	1.5
Sa	0	0	0	0.68	0	**99.32**	0
Su	0	0	0.9	0	0	0	**99.1**

2.4 Evaluation on AFEW database

In terms of recognition performance, AFEW is one of the most challenging video database. The EmotiW challenge shares the same data from AFEW. First, the authors evaluate of ECNN-FSA with different parameters on AFEW database and the results are shown in Table 6. From the first two lines of Table 4, the recognition accuracy of using one FC layer is 1.06% higher than that of using two FC layers. From lines 2 – 6, the recognition accuracy is influenced by super parameters and the highest recognition accuracy is achieved at s = 0.1. Furthermore, it can be seen from lines 2 – 7 that increasing the backbone output dimension will reduce the recognition performance. So, the output dimensions of backbone network and the enhanced branch are set to 1024, and the super parameter s is set to 0.1 in the later experiments.

Table 6 Evaluation of ECNN-FSA with different parameters on AFEW database

Parameters				
Output dimension of backbone network	Output dimension of enhanced branch	Number of FC layers	Super parameter s	Accuracy/%
1024	1024	2	0.06	40.91
1024	1024	1	0.06	41.97
1024	1024	1	0.01	41.14
1024	1024	1	0.1	**49.78**
1024	1024	1	0.2	49.25
1024	1024	1	0.3	43.77
2048	1024	1	0.06	40.64
2048	1024	2	0.06	39.76

Then, experiments are performed on different models and the parameters are adjusted on different model types. The experimental results are shown in Table 7. Here the meaning of the model type and parameters in Table 7 is explained. For example, 'ECNN-SA (3072, FC, s = 0.06)' means the model outputs 3072 dimensional feature vectors (the backbone CNN model outputs 2048-dimension and the enhanced branch outputs 1024-dimension), one FC layer and the super parameter s = 0.06. It can be seen that the output of GAP layer has the same number of channels as the input feature mapping, and the model increases the channel number by 1 × 1 convolution layer. Moreover, just extending feature mapping channels has no effect on the recognition results. Due to the strong fitting ability of FC layer, the model with one FC layer has better performance than the model with two FC layers. Furthermore, two hyper-parameters n and s of the proposed ECNN-FSA are evaluated to validate the robustness of the proposed method. For parameter n, besides the default value 5, several other values are tried, i.e. {3, 6, 9}, and the experimental results show that the performance is not sensitive to parameter n. For parameter s, based on 2048 dimensional features output by ECNN module and single FC layer model, the authors try several values, i.e. {0.001, 0.06, 0.1, 0.2, 0.3}. It can be seen that ECNN-FSA(2048, fc, s = 0.1) achieves the highest accuracy of 49.78%.

Table 7 Evaluation of ECNN-FSA with different parameters on AFEW database

	Model type	Accuracy
Proposed model	AFEW baseline	38.81%
	ECNN-LSTM(FC6, 7 × 7, LBP)	42.62%
	ECNN-LSTM(FC6, 7 × 7)	41.25%
	ECNN-FSA(3072, fc, s = 0.06)	41.64%
	ECNN-FSA(3072, 2 × fc, s = 0.06)	41.76%
	ECNN-FSA(2048, 2 × fc, s = 0.06)	42.91%
	ECNN-FSA(2048, fc, s = 0.06)	43.97%
	ECNN-FSA(2048, fc, s = 0.001)	43.14%
	ECNN-FSA(2048, **fc**, s = 0.1)	**49.78%**
	ECNN-FSA(2048, fc, s = 0.2)	49.25%
	ECNN-FSA(2048, fc, s = 0.3)	43.77%

Table 8 shows the comparison of the proposed ECNN-FSA with traditional CNN-LSTM in terms of end-to-end training and testing time of one frame. The training time of ECNN-FSA is reduced from 40.34 ms to 21.25 ms. Table 9 shows the performance comparison

of ECNN-FSA with CNN-LSTM in term of accuracy. The recognition accuracy of ECNN-FSA is 4.21% higher than that of CNN-LSTM.

Table 8 Training and testing time of ECNN-FSA and CNN-LSTM (ms)

CNN-LSTM[30]		Proposed ECNN-FSA	
Training time	Testing time	Training time	Testing time
40.34	8.26	21.25	5.57

Table 9 Performance of ECNN-FSA and CNN-LSTM on AFEW database

Model	Accuracy/%
Baseline	38.81
CNN-LSTM[30] (end-to-end)	41.57
Proposed ECNN-FSA (end-to-end)	49.78

Table 10 shows the confusion matrix of ECNN-FSA network on AFEW database. As it can be seen that the number of correctly classified test samples is not absolutely dominant, and sometimes the number of correctly classified videos is far less than the number of incorrectly classified videos. For example, only 13.64% of the videos labelled 'scared' are correctly classified, while 27.27% are wrongly classified as 'angry' and 15.91% are wrongly classified as 'happy'. Similarly, only 15.56% of the videos labelled 'surprise' are correctly classified, while 28.89% are wrongly classified as 'angry'. The reason is that human's mood is a mixture of multiple emotions, such as anger, disgust and sadness are usually associated with each other, and the facial morphology changes of fear, surprise and happiness have some similarities. Moreover, because the unconstrained facial expression data are mixed and interfered by many factors such as age, gender, race, illumination condition, posture change, occlusion, resolution, complex background, etc. So, it is a challenge to improve the video emotion recognition accuracy.

Table 10 Confusion matrix of ECNN-FSA network on AFEW (%)

	An	Di	Fe	Ha	Ne	Sa	Su
An	**50.0**	3.12	6.25	14.06	3.12	14.06	9.38
Di	7.50	**25.0**	2.5	22.5	17.5	15.0	10.0
Fe	27.27	4.55	**13.64**	15.91	6.82	25.0	6.82
Ha	4.84	1.61	0	**79.03**	4.84	6.45	3.23
Ne	10.17	8.47	8.47	20.34	**42.37**	6.78	3.39
Sa	5.0	8.33	3.33	20.0	6.67	**51.67**	5.0
Su	28.89	4.44	2.22	15.56	17.78	15.56	**15.56**

3 Conclusion

In this paper, a frame-level self-attention network for video-based facial expression recognition is proposed, i.e., ECNN-FSA. The ECNN-FSA is composed of a feature extraction module, a frame-level self-attention module, and a regression module. The feature extraction module adopts an improved VGG-16 network for abundant facial features, and the frame-level attention module makes the proposed model focus on key frames and capture the hierarchical structure of the video sequence at low computational cost. The experiments on CK+ and AFEW show that the proposed method can automatically capture the importance of frames, and the experimental results show that ECNN-FSA method outperforms other CNN based methods on AFEW database and obtains state-of-the-art results on CK+ database.

References

[1] MENG D B, PENG X J, WANG K, et al. Frame attention networks for facial expression recognition in videos[C]// Proceedings of IEEE International Conference on Image Processing (ICIP). Taibei: IEEE, 2019: 3866-3870.

[2] GHIMIRE D, JEONG S, LEE J. Facial expression recognition based on local region specific features and support vector machines [J]. Multimedia Tools and Applications, 2017, 76(6): 7803-7821.

[3] YANG H, CIFTCI U, YIN L. Facial expression recognition by de-expression residue learning[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City: IEEE, 2018: 2168-2177.

[4] FAN X J, TARDI T. A spatial-temporal framework based on histogram of gradients and optical flow for facial expression recognition in video sequences[J]. Pattern Recognition, 2015, 48(11): 3407-3416.

[5] RAHUL M, MAMORIA P, KOHLI N, et al. An efficient technique for facial expression recognition using multistage hidden Markov model[J]. Soft Computing: Theories and Applications, 2019, 742: 33-43.

[6] MIN H, HWA B, XWA B, et al. Video facial emotion recognition based on local enhanced motion history image and CNN-CTSLSTM networks[J]. Journal of Visual Communication and Image Representation, 2019, 59: 176-185.

[7] GUO C, LIANG J, ZHAN G, et al. Extended local binary patterns for efficient and robust spontaneous facial micro-expression recognition [J]. IEEE Access, 2019, 7: 174517-174530.

[8] WAN M H, LAI Z H, MING Z, et al. An improve face representation and recognition method based on graph regularized non-negative matrix factorization [J]. Multimedia

Tools and Applications, 2019, 78(15):22109-22126.
[9] GU J, HU H, XIE S. Enhanced dictionary pair learning sparse representation model for facial expression classification[C]// Proceedings of IEEE International Conference on Image Processing (ICIP). Beijing: IEEE, 2017: 4467-4471.
[10] GOODFELLOW I J, ERHAN D, CARRIER P L, et al. Challenges in representation learning: a report on three machine learning contests[C]// Proceedings of International Conference on Neural Information Processing. Daegu: ICML, 2013:117-124.
[11] DHALL A, GOECKE R, GHOSH S, et al. From individual to group-level emotion recognition: Emotiw 5.0 [C]// Proceedings of the 19th ACM International Conference on Multimodal Interaction. Glasgow: Association for Computing Machinery, 2017:524-528.
[12] LI Y J, ZHANG T. Deep neural mapping support vector machines [J]. Neural Networks, 2017, 93:185-194.
[13] JUNG H, LEE S, YIM J, et al. Joint fine-tuning in deep neural networks for facial expression recognition[C]// Proceedings of International Conference on Computer Vision (ICCV). Santiago: IEEE, 2015:2983-2991.
[14] MUNDHER A S, CHEAH W P, CONNIE T. Facial expression recognition using a hybrid CNN-SIFT aggregator [C]// Proceedings of International Workshop on Multidisciplinary Trends in Artificial Intelligence. Bandar Seri Begawan: Springer, 2017:496-499.
[15] TONG Y, CHEN R. Local dominant directional symmetrical coding patterns for facial expression recognition[J]. Computational Intelligence and Neuroscience, 2019(2): 231-243.
[16] TONG Y, CHEN R, LIANG R Y. Unconstrained facial expression recognition based on feature enhanced CNN and cross-layer LSTM[J]. IEICE Transactions on Information and Systems, 2020(11):1-4.
[17] CHEN R, TONG Y, LIANG R Y. Real-time generic object tracking via recurrent regression network [J]. IEICE Transactions on Information and Systems, 2020(3):602-611.
[18] DU T, LUBOMIR B, ROB F, et al. Learning spatiotemporal features with 3d convolutional networks[C]// Proceedings of International Conference on Computer Vision (ICCV). Santiago: IEEE, 2015:4489-4497.
[19] JAIN D K, ZHANG Z, HUANG K. Multi angle optimal pattern-based deep learning for automatic facial expression recognition[J]. Pattern Recognition Letters, 2017, 139: 230-232.
[20] YU Z, LIU Q, LIU G. Deeper cascaded peak-piloted network for weak expression recognition [J]. The Visual Computer, 2017, 34(12):1-9.
[21] IRSOY O, CARDIE C. Opinion mining with deep recurrent neural networks[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing. Doha: EMNLP, 2014:720-728.
[22] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural network[C]// Proceedings of International Conference on Neural Information Processing Systems. Kuching: MIT Press, 2014:3104-3112.
[23] SAMIRA E K, CHRISTOPHER P, XAVIER B, et al. Combining modality specific deep neural networks for emotion recognition in video[C]// Proceedings of the 15th ACM on International Conference on Multimodal Interaction (ICMI). Sydney: Association for Computing Machinery, 2013:543-550.
[24] KAHOU S E, BOUTHILLIER X, LAMBLIN P, et al. Emonets: multimodal deep learning approaches for emotion recognition in video[J]. Journal on Multimodal User Interfaces, 2016, 10(2):99-111.
[25] TZIRAKIS P, TRIGEORGIS G, NICOLAOU M A, et al. End-to-end multimodal emotion recognition using deep neural networks[J]. IEEE Journal of Selected Topics in Signal Processing, 2017, 11(8):1301-1309.
[26] SARAH A B, EMAD B, CRISTIAN C F, et al. Emotion recognition in the wild from videos using images[C]// Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI). Tokyo: Association for Computing Machinery, 2016:433-436.
[27] XU K, BA J L, KIROS R, et al. Show, attend and tell: neural image caption generation with visual attention[C]// Proceedings of International Conference on Machine Learning. Lille: Association for Computing Machinery, 2015:2048-2057.
[28] LIU M, SHAN S, WANG R, et al. Learning expression lets on spatiotemporal manifold for dynamic facial expression recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Columbus: IEEE, 2014:1749-1756.
[29] VASWANI A, SHAZEER N, PARMAR N. Attention is all you need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2017:6000-6010.
[30] FAJTL J, SOKEH H S, ARGYRIOU V. Summarizing videos with attention[C]// Proceedings of Asian Conference on Computer Vision (ACCV). Perth: Springer, 2018, 541:39-54.
[31] LUCEY P, COHN J F, KANADE T, et al. The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression[C]// Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco: IEEE, 2010:94-101.
[32] DHALL A, GOECKE R, LUCEY S, et al. Collecting large, richly annotated facial-expression databases from movies [J]. IEEE Multimedia, 2012, 19(3):34-41.
[33] DHALL A, GOECKE R, GHOSH S, et al. From individual to group-level emotion recognition: Emotiw 5.0[C]//

Proceedings of the 19th ACM International Conference on Multimodal Interaction. Glasgow: Association for Computing Machinery, 2017:524-528.

[34] SIKKA K, SHARMA G, BARTLETT M. LOMO: latent ordinal model for facial analysis in videos[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016:5580-5589.

CHER Rui, born in 1972. She received her Ph. D degree from Nanjing University of Post and Telecommunications in 2013. She received her B. S. and M. S. degrees from Southeast University, China, in 1991 and 1996 respectively. She is mainly engaged in the research of object detection and tracking.