IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 15, 2019, accepted May 4, 2019, date of publication May 10, 2019, date of current version May 28, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2916211

Learning Hierarchical Emotion Context for Continuous Dimensional Emotion Recognition From Video Sequences

QIRONG MAO, (Member, IEEE), QING ZHU, QIYU RAO, HONGJIE JIA, AND SIDIAN LUO
[1]School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang 212013, China

Corresponding author: Qirong Mao (mao_qr@ujs.edu.cn)

This work was supported in part by the National Nature Science Foundation of China under Grant 61672267, in part by the Key Projects of the National Natural Science Foundation of China under Grant U1836220, and in part by the Innovation Project of Undergraduate Students in Jiangsu Province under Grant 201710299088X.

ABSTRACT Dimensional emotion recognition is currently one of the most challenging tasks in the field of affective computing. In this paper, a novel three-stage method is proposed to learn hierarchical emotion context information (feature- and label-level contexts) for predicting affective dimension values from video sequences. In the first stage, a feed-forward neural network is used to generate a high-level representation of the raw input features. Then, in the second stage, the bidirectional long short-term memory (BLSTM) layers learn the context information of the feature sequences from the high-level representation and get the initial recognition results of the input. Finally, in the third stage, a BLSTM neural network is used to learn the context information from emotion label sequences by an unsupervised way, which is used to correct the initial recognition results and get the final results. We also explore the influence of different sequence lengths by sampling from the original sequences. The experiment performed on the video data of AVEC 2015 demonstrates the effectiveness of the proposed method. Our framework highlights that incorporating both feature/label level dependencies and context information is a promising research direction for predicting the continuous dimensional emotion.

INDEX TERMS Affective computing, emotion dimension, hierarchical emotion context learning, video expression, BLSTM.

I. INTRODUCTION

Emotional states play a fundamental and important role in human communication. Understanding human emotional states are important for human-human interaction and social contact. Hence automatic emotional state recognition has been an active research area in the past years [1]–[3].

According to theories in psychology research [4], [5], there are three emotion theories to model the emotion state: discrete theory, appraisal theory and dimensional theory. The discrete theory claims that there exists a small number of discrete emotions (i.e., angry, disgust, happiness, neutral, sadness, afraid, and surprise) that are basic in our brain and recognized universally [6]. In research on automatic emotional state recognition, this intuitive and simple theory interpreting emotional states as basic categories has been the most commonly adopted approach. However, people exhibit non-basic, subtle and complex emotional states like depression. Therefore, basic discrete classes may not reflect the complexity of the emotional state expressed by human. Hence, many researchers advocate the use of dimensional theory. In the appraisal theory, emotions are generated through continuous, recursive subjective evaluation of both our own internal state and the state of the outside world [6]. However, this theory is still an open research problem on how to use it for automatic measurement of emotional state. Dimensional theory claims an emotional state as a point in a continuous space. Hence, this dimensional theory can model the subtle, complicated and continuous emotional state.

In recent years, there has been a shift towards predicting emotion in continuous dimensional space from recognition of discrete emotion categories. Many researches have been investigated on video sequences to get continuous recognition in the dimensional space [7]–[10]. Typically, valence (V) and

The associate editor coordinating the review of this manuscript and approving it for publication was Yonghong Peng.

arousal (A) dimensional space is one of the most popular continuous emotional model. The valence dimension refers to how positive or negative the emotion is, and range from unpleasant to pleasant. The arousal dimension refers to how excited or apathetic the emotion is, and it ranges from sleepiness or boredom to frantic excitement [11].

For continuous dimensional emotion recognition, the challenge is to build systems that can continuously (i.e., over time) analyze and predict affective emotion in dimensional space. The survey by Sariyanidi *et al.* [12] highlighted the importance to make better use of the context information on sequential data. Context information is very useful because the target dimensional values are continuous and have a short time gap between two adjacent predictions. In a video sequence, an emotional expression of a video frame is recognized by taking into account not only the input feature describing that frame, but also the input features describing the previous frames and the future frames, i.e., how the expression evolved over time to the current state, how the expression evolve from current state to future state [8].

Some previous work which are able to make use of context information, but they are very tied to the feature level. As suggested in [12], there is a significant gap between feature level and semantic emotion level (label level) in sequential data. For example, face images can change fast and dramatically in videos, even if the emotional state of the person will change at a slower speed [8]. And there are much irrelevant information (e.g., head pose, illumination) at the feature level. When those irrelevant information change fast and dramatically, the context information between the feature sequence can be very destructive to emotional recognition. And there is another problem between feature level and label level which is that the emotional values are annotated by human person. When annotators make the decision based on vocal and visual signals, there is an inherent annotation delay between their observations and decisions. Thus, it is important to explore new strategies that can exploit the context information at both feature level and label level, and the same important as handling with the high variability of the feature sequence [13]. Therefore, it is necessary to carry out dimensional emotion recognition based on context information.

In this paper, we use the three-stage framework based on BLSTM, we tackle the problem of continuous dimensional emotion recognition by learning hierarchical emotion context information at feature level and label level. In the approach, context information at feature level is firstly used to do soft emotion recognition and get the initial emotion recognition result, which will be corrected by using the emotion context information of label level in the third stage. The major contributions of this paper are:

1) We proposed a model combined with feed-forward neural network and BLSTM networks to capture context information on the feature sequence. The feed-forward neural network is used to generate a high-level representation of the raw input features, and then BLSTM learns the context information of feature sequences from the high-level representation. It will reduce the affection of useless factors for emotion recognition.
2) A novel three-stage framework based on BLSTM is proposed to learn the hierarchical emotion context information, namely high-level feature representation learning, feature-level emotion context learning (FLECL) and label-level emotion context learning (LLECL). Specifically, we will get the initial recognition result at the second stage by making use of the feature-level context information. And at the third stage, the final recognition result is got by using the label-level emotional context information learned from the emotion label sequences.
3) We explore the influence of the sequence length by sampling different lengths of sequences from the original sequences, and then find the appropriate length for emotional context learning for continuous emotion recognition.

The rest of the paper is organized as follows. We introduce the related work in Section II. Section III presents our BLSTM-based hierarchical context learning algorithm in detail. Section IV describes datasets and reports our experimental results. Conclusions and future directions are discussed in Section V.

II. RELATED WORK

A. DIMENSIONAL EMOTION RECOGNITION

Dimensional space is able to represent a wide range of subtle and complicated emotions, especially those spontaneous non-prototypical ones in real-life data [10]. Various continuous dimensional emotion recognition systems have been built. The typical approach is to take every single data as a single unit (e.g., a frame of a video sequence) independently. It can be made as a standard regression problem for every frame using the so-called static (frame-based) regressors. Gunes and Pantic in paper [14] focus on dimensional recognition of emotions from head gestures using the Support Vector Machines for Regression (SVR). Another more interesting approach uses the temporal relationship between different continuous data to make a better recognition of emotions. In [29], authors use both CNNs and RNNs to set up a system that performs emotion recognition on video data. The system with the dimensional approaches can model truly the emotion over time so that can output time-continuous labels.

B. CONTEXT LEARNING FOR DIMENSIONAL EMOTION RECOGNITION

As discussed in section I, context information is very important for continuous dimensional emotion recognition. Many researchers explored techniques to take advantage of this information. In [15], authors use spatiotemporal representations by computing features over a temporal window rather than a single frame. But there is a common assumption that within a window of expression there are no head pose variations but only facial activity changes which are not common in real-life data [8], [12]. Another typical method makes use of context information is Hidden Markov Model (HMM). It has been applied for facial expression recognition [16], speech emotion recognition [17], and audio-visual affect recognition [18]. In the following, we will take a short review of the feature-level and label-level emotion context learning respectively.

1) FEATURE-LEVEL EMOTION CONTEXT LEARNING

Long Short Term Memory (LSTM) [19] has been successfully used for modeling the context information. LSTM is one type of recurrent neural network (RNN), has been proven to be effective for modeling the relationship between sequential observations by making use of past information. This method outperformed techniques such as support vector regression (SVR) [20]. Many researchers used it to learn the feature-level context information on audio/video data and the results showed its promising performance [20]–[22]. Wöllmer *et al.* [23] first proposed a method based on LSTM recurrent neural networks for continuous emotions recognition that includes modeling of long-range dependencies between observations. In [9], authors use Bidirectional Long Short Term Memory (BLSTM) to do affective dimension recognition. BLSTM can model long-range temporal dependencies on sequential observations by using past and future contexts. In this study, we chose an approach based on BLSTM recurrent neural network because it is capable of modeling time series with contextual dependencies.

2) LABEL-LEVEL EMOTION CONTEXT LEARNING

The problem of synchronization of the feature level and emotion level (label level) has been investigated in the literature. In [24], Hung et al. investigated annotation delay compensation by applying temporal shifts and smoothing filters. Temporal shifts is a way to realign the feature with the ground truth. Smoothing filter is to realign the recognition results. Many researches use the two techniques to deal with delay problems [9], [20], [24], but it is difficult to decide how many frames should be shifted or the length of the filter. In [9], authors used a deep BLSTM recurrent neural network to get initial recognition results and then they adopted two Gaussian smoothing methods: one with window of fixed length (120 frames), the other with windows of variable length.

In [8], [25]–[27], multistage approaches have been proposed to learn the information in different levels. Graves *et al.* [7] trained a two-stage system for classifying every individual video frame, one stage is a traditional regression method, and another is a time-delay neural network for temporal relationships between consecutive recognition results. Meng and Bianchi-Berthouze *et al.* [27] trained a multi-layer hybrid framework composed of a temporal regression layer for recognizing emotion dimensions, a graphical model layer for modeling valence-arousal correlations, and a final classification and fusion layer exploiting informative statistics extracted from the lower layers. In [25] and [27], a multi-stage approach was proposed to separate the feature level and the decision level. In the feature level, traditional classification methods were used to predict the emotion labels. In the decision level, the transitions (over time) between consecutive affective dimension levels were modeled as a first-order Markov model. The main limitation of current multistage approaches is that they do not make use of connections between feature level and label level. They used traditional classification methods (e.g., SVM, SVR, and KNN) in the feature level, and then analysis temporal information in the label level or decision level separately.

In this paper, our model to learn hierarchical emotion context information mainly based on BLSTM. There are several studies that make use of LSTM variants for dimensional emotion recognition [20]–[22], whereas these previous works mainly focus on learning the context information at the feature level. In this paper, we introduce a hierarchical context learning framework for continuous dimensional emotion recognition by using BLSTM, in which we deal with the problem of learning the context information at both feature level and label level through a novel three-stage framework. When learning the context information of feature level, instead of inputting the raw feature into BLSTM layer directly, we use the features learned by the feed-forward neural network first as the input of the BLSTM layers. And when learning the context information of label level, we use the emotion labels to train the BLSTM network in an unsupervised way. We also compare our model with several studies that make use of LSTM variants. And we also explore the influence of the different lengths of sequences by sampling. During sampling, we also explore the proper sampling step length according to the emotion duration.

III. LEARNING THE HIERARCHICAL CONTEXT AMONG FEATURES AND LABELS FOR CONTINUOUS DIMENSIONAL EMOTION RECOGNITION

In this section, we first describe the overall architecture, and then present the algorithms used in this paper.

A. SYSTEM ARCHITECTURE

Fig. 1 shows the architecture of the proposed continuous dimensional emotion recognition system, which has one hidden layer for the High-Level Feature Learning(HLFL), three BLSTM layers for Feature-Level Emotion Context Learning (FLECL) and one BLSTM layer for Label-Level Emotion Context Learning (LLECL). The input of our model is the appearance feature vectors extracted from each frame intercepted in video. We first use the low-level video feature sequences as the input of the hidden layer to learn the high-level representation of the input video. Then BLSTM is adapted to learn the feature-level emotional context from the high-level feature representation, and then give the initial emotional dimension label recognition. Finally, the initial emotional dimension label recognition results are used as the input of the BLSTM at third stage to be further corrected by using the label-level emotional context information, and get

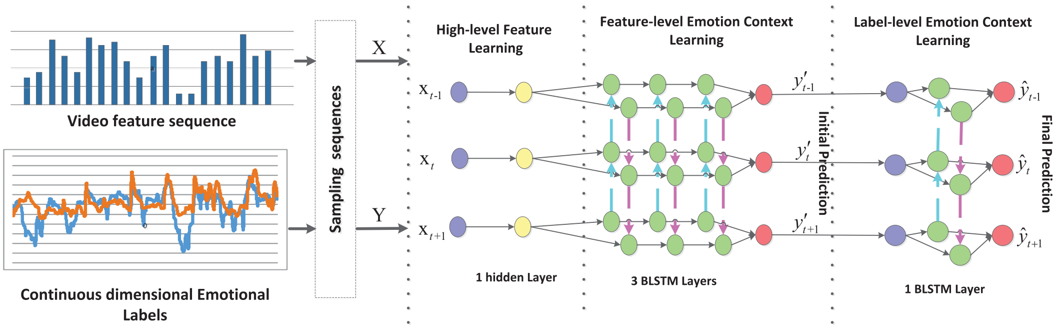


FIGURE 1. System architecture.

the final emotion recognition results. Specifically, the training can be broken into the following three stages:

- **High-level Feature Learning:** At this stage, the raw feature sequences are fed into the hidden layer to learn high-level representations since the high-level representation of the raw features can reduce the variability of feature sequences, which makes context learning more easily.
- **Feature-level Emotion Context Learning:** The high-level feature representations learned at the first stage are fed into the BLSTM network to the learn feature-level context. This stage and the first stage are trained together with a connected neural network by the supervised way. And the output of this stage is the initial recognition of emotion values.
- **Label-level Emotion Context Learning:** At this stage, we use BLSTM network to learn the emotional context of the label level. Specifically, we train this BLSTM network in an unsupervised way. The input and output are both ground truth emotion value sequence. We apply it to capture the changing pattern of emotion state sequences. The initial recognition emotion results of the second stage are used as the input of this stage to be corrected further. The output of this stage is the final recognition results.

B. HIGH-LEVEL FEATURE REPRESENTATION LEARNING

Instead of inputting the raw feature into BLSTM layer directly, we put the low-level feature into the feed-forward neural network first, and the output of which is fed into the BLSTM layers. We use the feed-forward neural network to learn a high-level feature from the input feature sequence. Given an input feature sequence $X = (x_1, x_2, \ldots, x_T)$, x_t is the t-th feature in that sequence, and x_t represents the input feature at time t. The output of this network is the high-level feature sequence $X' = (x'_1, x'_2, \ldots, x'_T)$. We use the sigmoid function as the activation function and the formula of one layer is as follows:

$$\begin{aligned} z_t &= Wx_t + b, \\ x'_t &= \frac{1}{1+e^{-z_t}}, \end{aligned} \tag{1}$$

where W is the weight matrix of that layer, b is the bias. In our experiments, we use more than one layer to learn the high-level feature. The feature dimension of x_t is 84. We set the layer with 64 hidden units, so the dimension of W is 84*64, and the feature dimension of the high-level feature x'_t is 64.

C. FEATURE-LEVEL EMOTION CONTEXT LEARNING

This stage focuses on the feature-level context learning (the video feature sequence) by using BLSTM. It performs a typical supervised training process with a neural network to produce an initial recognition for each video frame.

More specifically, a standard LSTM layer computes the hidden state sequences $h = (h_1, h_2, \ldots, h_T)$ by iterating the following equations from $t = 1$ to T.

$$\begin{aligned} i_t &= \sigma(W_{xi}x'_t + W_{hi}h_{t-1} + b_i) \\ f_t &= \sigma(W_{xf}x'_t + W_{hf}h_{t-1} + b_f) \\ c_t &= f_t * c_{t-1} + tanh(W_{xc}x'_t + W_{hc}h_{t-1} + b_c) \\ o_t &= \sigma(W_{xo}x'_t + W_{ho}h_{t-1} + b_o) \\ h_t &= o_t * tanh(c_t), \end{aligned} \tag{2}$$

where σ is the sigmoid function, i, f, o and c denote the *input gate, forget gate, output gate* and *cell activation* respectively, W is the weight matrices. Here, we use H to denote LSTM layer operation, which is

$$(h_t, c_t) = H(x'_t, h_{t-1}, c_{t-1}). \tag{3}$$

In a BLSTM layer, there are two directional operations, the past and the future time direction, which make it capture available context information in both the past and the future of a specific time frame in a sequence. We use the output x'_t of the hidden layer as the input of BLSTM network. Consider that there are N layers (without input layer) in the network and the length of an input sequence is T, then the operations of the network are as follows:

$$\begin{aligned} (\overrightarrow{h}^n_t, \overrightarrow{c}^n_t) &= \overrightarrow{H}^n(\overrightarrow{h}^{n-1}_t, \overrightarrow{h}^n_{t-1}, \overrightarrow{c}^n_{t-1}), \\ (\overleftarrow{h}^n_t, \overleftarrow{c}^n_t) &= \overleftarrow{H}^n(\overleftarrow{h}^{n-1}_t, \overleftarrow{h}^n_{t-1}, \overleftarrow{c}^n_{t-1}), \end{aligned} \tag{4}$$

$$y'_t = W_{\overrightarrow{h}^{N-1}y}\overrightarrow{h}^{N-1}_t + W_{\overleftarrow{h}^{N-1}y}\overleftarrow{h}^{N-1}_t + b_y, \tag{5}$$

where $t = 1$ to T, $n = 2$ to $N-1$, $\overrightarrow{h}_t^1 = x'_t$, $\overleftarrow{h}_t^1 = x'_t$, and x_t the input, y'_t the output. $\overrightarrow{h}_0^n$, $\overleftarrow{h}_0^n$, $\overrightarrow{c}_0^n$ and $\overleftarrow{c}_0^n$ are randomly initialized for all BLSTM layers. Equation(1) is the operation of the multi-layer perceptron. Equation(4) is the operation of the multi BLSTM layer. Notice here in (4), when the hidden state feed forward into the next layer, it only feeds into the layer which has the same direction (positive or negative time direction) with current layer. Equation(5) is the operation of the linear regression layer. The output is calculated by taking into account of both directions (positive and negative time, or, past and future).All the operations in (1), (4), and (5) are done iteratively from $t = 1$ to T.

Assume y' denotes the estimation result of the dimensional emotion value y and T is length of sequence, then Root Mean Square Error (RMSE) is defined as:

$$RMSE = \sqrt{\frac{1}{T}\sum_{1}^{T}(y_t - y'_t)^2}. \tag{6}$$

We choose the RMSE as the loss function of the feature-level emotion context learning stage and the label-level emotion context learning stage.

By combining the feed-forward neural network layer and the BLSTM layers, the model will firstly use the regression layer to generate a high-level representation of the raw input features, then the BLSTM learn the context information from the high-level representation. When the basic hidden layer's outputs are fed into BLSTM layer, this method will reduce the feature dimension which makes BLSTM's computation become easy.

D. LABEL-LEVEL EMOTION CONTEXT LEARNING

This stage focuses on the label-level emotion context learning. After the second stage, we can get the initial predicted dimensional emotion sequence $Y' = (y'_1, y'_2, \ldots, y'_T)$ from feature sequence. We also have the ground truth label sequence $Y = (y_1, y_2, \ldots, y_T)$. As discussed above, there is a gap between the context of the feature level and the label level(Y' and Y). We invested our work in overcoming this problem.

In order to capture the label-level emotional context information from the continuous emotion label sequence, in this stage of training, we use the continuous emotion label sequence as the input to train the BLSTM network with the capacity of reconstructing the given input sequence. The input and output both are the ground truth of the emotional value. The whole this training stage is similar to an autoencoder, we regard the ground truth of the emotion value as the "unlabeled features" to learn the emotional context information, hence, we consider the whole process is in an unsupervised way. The ground truth of the emotion value is used as the input feature of the BLSTM network by reconstructing to learn the label-level emotional context information from the ground truth of emotion label sequence. Our purpose is to use the training method of regression to learn the context information of label-level. In our experiments, a multi-layer architecture did not improve the performance. Hence a BLSTM layer followed by a linear regression layer is adapted to reconstruct the input sequence. We use $\hat{Y} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_T)$ to denote the output and the RMSE between $\hat{Y}$ and Y as the loss function of this stage. The main purpose of this stage is to let the output $\hat{Y}$ get close to the ground truth of emotion labels Y. Thus, our model can learn the label-level emotional context information. When testing, our purpose is using this label-level emotional context to correct the initial recognition results Y' got in the second stage. Then we take the output of the third stage as the final emotion recognition results.

IV. EXPERIMENTS & RESULTS

A. DATA SET & EXPERIMENTAL SETUP

We use the dataset provided by the AVEC2015 challenge [20] to evaluate the performance of our method. This dataset is a subset of Remote Collaboration and Affective Interaction (RECOLA) [28]. Spontaneous and naturalistic interactions were collected during the resolution of a collaborative task through video conference. The dataset is annotated in two emotion dimension, arousal and valence, by 6 French speakers in scales $[-1, 1]$ for every 40ms. Data was recorded by audio, video, electro-cardiogram (ECG) and electro-derma (EDA) modalities. Due to our research interest, we only use the information on the video. Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) using in AVEC challenge is used as the appearance features of our experiment. Then Principal Component Analysis is performed on the LGBP-TOP, resulting in 84 dimensional feature vectors with the frame rate of 25 frames/s. The dataset is equally divided into three partitions: training set, development set and testing set, each having 9 recordings of 5 minutes.

Since we do not have the ground truth labels of the test set, we sampled sequences from the training set as our training samples, the original training set is used as the validation set, and the original development set is used as our testing set. Therefore, the experiment results in this paper are reported on the original development set which we use as the testing set.

We evaluate our hierarchical emotion context learning method based on the RMSE and CCC, and compare the performance of our model in different cases: 1) with and without label-level emotion context learning and 2)with and without feed-forward neural network layer in feature-level emotion context learning. Finally, we give intuitive recognition examples compared with the ground truth for three cases of our system: 1) with and without the high-level feature learning stage; 2) feature-level emotion context learning trained with original sequences and sampled sequences; 3) with and without label-level emotion context learning.

To further evaluate the performance of the proposed method, we compared our method with other well-established methods in paper [9], [20]–[22], [29]. The baseline 1 in paper [20] used a hybrid decision-fusion network based on

Support Vector Regression (SVR) and Neural Network. For SVR, they used a linear kernel. For the neural network, they used three setups: feed-forward, LSTM and BLSTM. And they applied a median-filtering with window size in $[0.2-20]$s. Baseline 2 provides the CCC metric obtained by LSTM neural network which got the best performance in their experiments, and the LSTM network has two layers with 90 and 60 hidden units respectively. The work of Chao *et al.* [21] used LSTM layers combined with hidden layer and temporal pooling layer to smooth the feature sequence. The system has two LSTM layers and each layer has 64 hidden units, and they used the mean pooling in temporal pooling layer. Chen *et al.* [22] used a two-hidden layer LSTM model and they adopt MSE as the loss function. He *et al.* [9] used a Deep Bidirectional Long Short-Term Memory recurrent neural networks (DBLSTM) to get the initial emotion recognition results, and adopt Gaussian smoothing method with the moving window of fixed length to smooth the initial recognition results. Before features are input into the DBLSTM model, they offset the feature sequence and the label sequence. Khorrami *et al.* [29] used two models to recognize valence emotion: single frame CNN model with dropout and CNN + RNN model. The CNN model they used has 3 convolutional layers consisting of 64, 128, and 256 filters respectively, the size of which is 5 × 5. The first two convolutional layers are followed by 2 × 2 max pooling whereas the third layer is followed by quadrant pooling. The convolutional layer is a fully-connected layer with 300 hidden units, and a linear regression layer is used to estimate the valence label. The RNN model they used has a single layer RNN with 100 units in the hidden layer.

We implemented our model based on Theano [30]. The model is trained with Adadelta optimization algorithm. In order to speed up the convergence and increase the generalization ability, we adopt an early stop strategy. The parameters of the machine the experiments used are Intel Core i7-4790 CPU, 16G RAM and NVIDA GTX 780Ti.

B. EVALUATION METRICS

The performance is reported in terms of RMSE (shown in Equation 6) and Concordance Correlation Coeffcient (CCC). CCC is defined as:

$$CCC = \frac{2\rho\delta_y\delta_{y'}}{\delta_y^2 + \delta_{y'}^2 + (\mu_y - \mu_{y'})^2} \tag{7}$$

where ρ is the Person correlation coefficient (CC) between two time series (*e.g.*, recognition and ground truth), and δ_y and $\delta_{y'}$ are the variance of every time series. μ_y and $\mu_{y'}$ are the mean value of each time series. Since it combines the CC and the mean square error (MSE), CCC is more reliable for the evaluation of the regression problem. Therefore, recognition well correlated with the ground truth but shifted in value are penalized in proportion to the deviation.

From the definition of RMSE and CCC, we can see that the model with smaller RMSE and higher CCC has better performance.

C. PARAMETER SELECTION

1) PARAMETER SELECTION OF NETWORK

We conducted experiments with the number of feed-forward neural network layer and BLSTM layer range from 1 to 8 with step 1, and the number of hidden units range from $0.25 \times M$ to $1.25 \times M$ with step $0.25 \times M$ respectively on the validation sequences, where M is the dimension of the input feature. The parameters that give the best results on the validation set are chosen as the final parameters. Finally, our model has 1 feed-forward neural network layer, 3 BLSTM layers for feature-level context learning, 1 BLSTM layers for label-level emotion context learning and 1 linear regression layer for recognition. The number of hidden units for each layer is chosen as [64, 64, 64, 64, 1].

2) SEQUENCES SAMPLING

In this section, we conduct the experiment to explore the best length of the video sequence by sampling different length of sequences. Assume that the length of the original sequence is T, the length of the sampled sequence is T', and $T' = \alpha \times T$. We use a step β to sample the sequence. Specifically, the first sampled sequence is $[0, \ldots, 0 + T']$, the second sequence is $[\beta, \ldots, \beta + T']$. By doing so, we can get n sequences from one original sequence, where $n \times \beta <= T - T'$ and $(n+1) \times \beta > T - T'$. In addition, by using this sampling method, we will get more training data to train the model.

In our experiments, we conduct experiments with value of α ranges from 0.5 to 0.95 with step 0.05 and β is in [100, 125, 150]. The overall duration of emotion is supposed to fall between 0.5 and 4 seconds. When we sampled the short sequence with step β we want every sampled sequence to begin with a new emotion. The features are extracted every 40ms, and we sample sequence every 4s at least to make sure that every sampled sequence may begin with a new emotion. Hence, the value of β is large than 100 (4s/40ms).

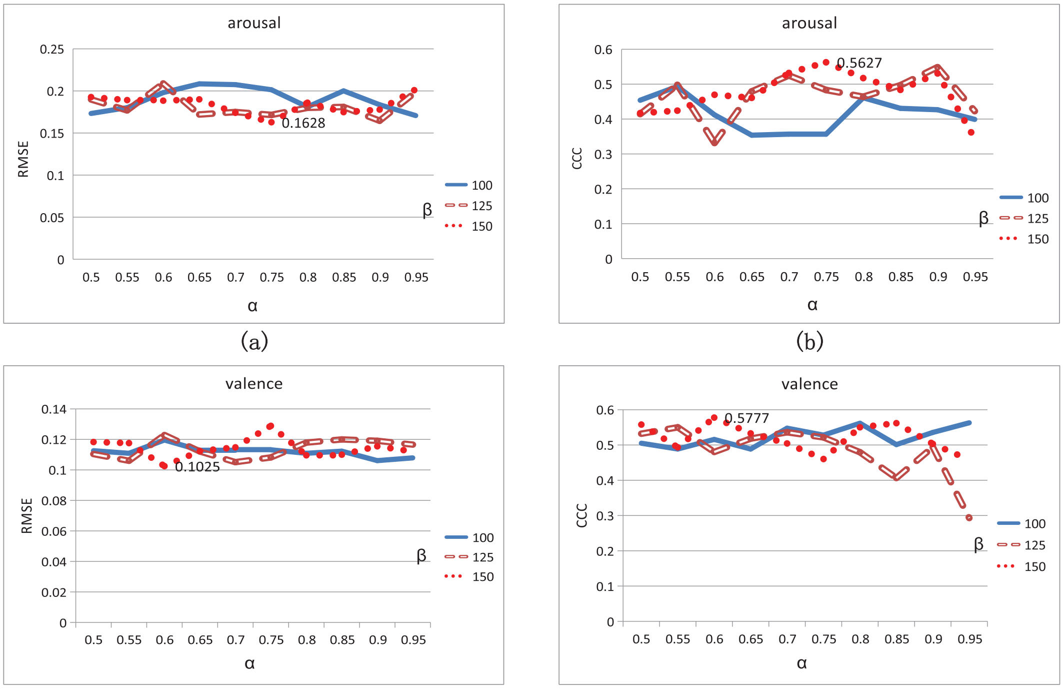
The results at the FLECL stage are shown in Fig. 2. We can see that the sequences with a large step value (β) can get better performance. When $\alpha = 0.75$ and $\beta = 150$, the CCC is 0.5627 and the RMSE is 0.1628 on the arousal dimension and when $\alpha = 0.6$ and $\beta = 150$, the CCC is 0.5777 and the RMSE is 0.1025 in the valence dimension. Thus, we fix α at 0.75 in the arousal dimension, α at 0.6 in the valence dimension and β at 150 in both dimension at the FLECL stage.

We also conduct the experiments with the same configuration at the LLECL stage. When α is 0.95 and β is 125, the best performance is obtained in both arousal and valence dimension at the LLECL stage. Therefore, we fix α as 0.95 and β as 125 in the LLECL stage in the following experiments.

D. PERFORMANCE EVALUATION

1) COMPARISON WITH THE METHODS WITH AND WITHOUT HIGH-LEVEL FEATURE LEARNING

In order to evaluate the performance of the high-level feature learning(HLFL) stage in our method, we compare our



(a)

(b)

(c)

(d)

FIGURE 2. Performance with different α and β in arousal and valence recognition. (a) RMSE on arousal with different α and β. (b) CCC on arousal with different α and β. (c) RMSE on valence with different α and β. (d) CCC on valence with different α and β.

TABLE 1. Comparison results of the methods with/without high-level feature learning. The highest CCC and the smallest RMSE and training time are highlighted in bold.

Layers	Arousal			Valence		
	RMSE	CCC	TIME(s)	RMSE	CCC	TIME(s)
without HLFL	0.2153	0.1673	53.2815	0.1644	0.1480	53.2328
with HLFL	**0.1817**	**0.4179**	**43.4851**	**0.1256**	**0.3095**	**43.9417**

TABLE 2. Comparison results of the methods with and without label-level emotion context learning. The highest CCC and the smallest RMSE is highlighted in bold.

	Arousal		Valence	
	RMSE	CCC	RMSE	CCC
FLECL	**0.1628**	0.5627	**0.1025**	0.5777
FLECL+LLECL	0.1654	**0.5797**	0.1036	**0.5954**

proposed model with the network without HLFL stage. The model without HLFL has 4 BLSTM layers and our model with HLFL has 1 hidden layer 3 BLSTM layer. The comparison results are shown in Table 1. From Table 1, we can see that the proposed network has a significant improvement on RMSE and CCC in both arousal and valence dimension. We also compare the training time of one epoch of these two methods. We can see that the proposed method takes less training time compared with the basic 4 BLSTM layers network. The result shows that by introducing the high-level feature learning stage, the model can learn the feature context much better and easier.

2) COMPARISON WITH THE METHODS WITH AND WITHOUT LABEL-LEVEL EMOTIONAL CONTEXT LEARNING

In order to evaluate the performance of label-level emotional context learning, we conduct the experiments by using the methods with and without the LLECL stage. The results are listed in Table 2. Table 2 clearly shows that the proposed method in this paper outperforms the model without LLECL stage, and gets an obvious increase for the CCC value and a little increase for the RMSE value. As mentioned above, the model with smaller RMSE and higher CCC has better performance. Our model gets an obvious increase in performance for the CCC metric and a litter decrease for the RMSE metric. CCC combines the CC and the mean square error (MSE), it is much more reliable for evaluating the regression problem. Hence, the LLECL stage can improve the performance of our method.

3) COMPARISON WITH OTHER WELL-ESTABLISHED METHODS

To further evaluate the performance of the proposed method, we compared our method with the other seven well-established methods early discussed in the third paragraph of Section IV.A. The comparison results are listed in Table 3. Except for the method in paper [29], the feature set used in this comparison experiment is the video appearance feature set provided by AVEC2015. The method in paper [29] used their own feature set, and the authors

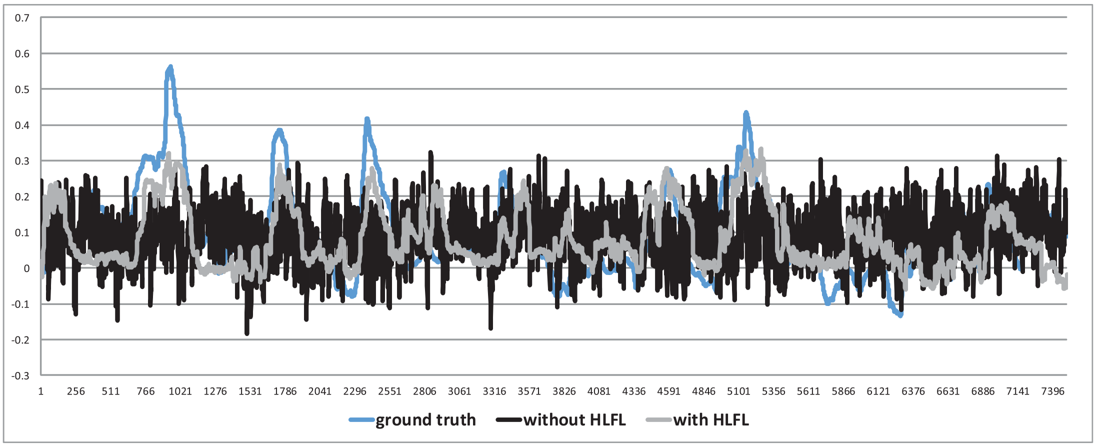


FIGURE 3. Valence recognition: With and without high-level feature learning stage.

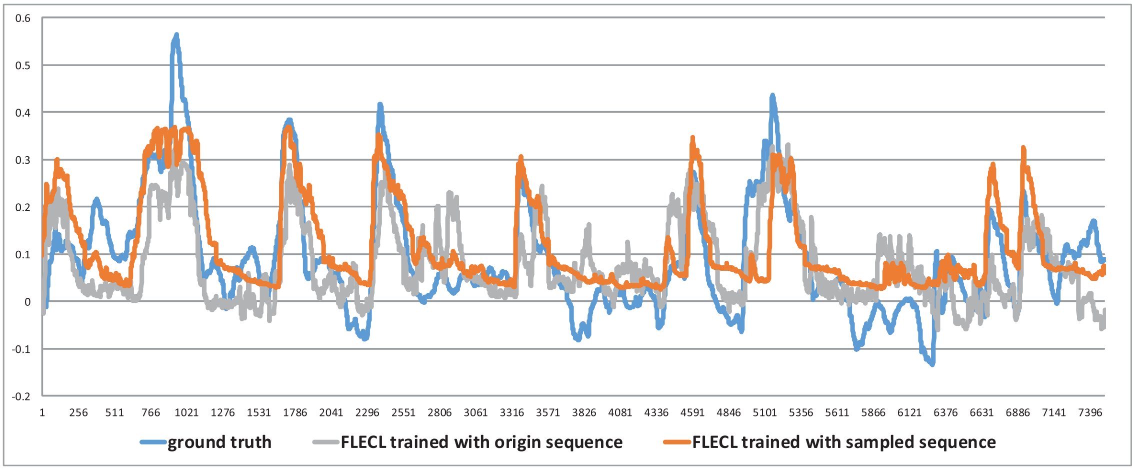


FIGURE 4. Valence recognition: FLECL trained with original sequences and the sampled sequences respectively.

also compared the same models as we compared with. In order to clearly describe the compared methods in Table 3, we use the logogram to replace the whole name of models in paper [29]. The detailed information is single frame CNN model with dropout(Khorrami et al. 1) and CNN + RNN model(Khorrami et al. 2). Therefore, we use the experiment results listed in paper [29] directly to compare. The results reported in Table 3 are all on the development set provided by AVEC2015. From the comparison, we can see that our model gets the superior/highly-competitive performance on both RMSE and CCC metrics when compared with the other well-established methods.

TABLE 3. Comparison with the other well-established methods. The highest CCC and the smallest RMSE are highlighted in bold. The CCC of the arousal and the valence for each method are averaged in the last column.

Method	Arousal		Valence		Avg.
	RMSE	CCC	RMSE	CCC	CCC
Baseline 1 [20]	0.214	0.103	0.117	0.273	0.188
Baseline 2 [20]	-	0.079	-	0.273	0.176
Linlin Chao et al. [21]	0.188	0.535	0.121	0.463	0.499
Shizhe Chen et al. [22]	0.168	0.533	0.114	0.354	0.444
Lang He et al. [9]	0.185	0.226	0.105	0.346	0.286
Khorrami et al. 1 [29]	-	-	0.114	0.363	-
Khorrami et al. 2 [29]	-	-	0.106	0.489	-
our model	**0.165**	**0.580**	**0.104**	**0.595**	**0.588**

4) RECOGNITION VISUALIZATION

In order to show the performance of our method intuitively, in this section, we provide three examples of valence dimension recognition on the dev_9 video file respectively for the comparison of FLECL with and without high-level feature learning, FLECL trained with original sequence and sampled sequences, the methods with and without LLECL stage.

Fig. 3 provides the recognition results made by FLECL and without high-level feature learning stage. From Fig. 3, it can be seen that the FLECL with HLFL has a better recognition result compared with the method without HLFL. Moreover, the recognition of the method with HLFL is more smooth than the method without HLFL, and is more close to the ground truth labels. Since the raw input feature sequence may have a high variability which is bad for BLSTM. By learning a high-level feature, it can reduce the feature variability.

Fig. 4 provides the comparison results of the recognition made by the FLECL stage trained with original sequence and sampled sequences. Fig. 4 clearly shows that, with the

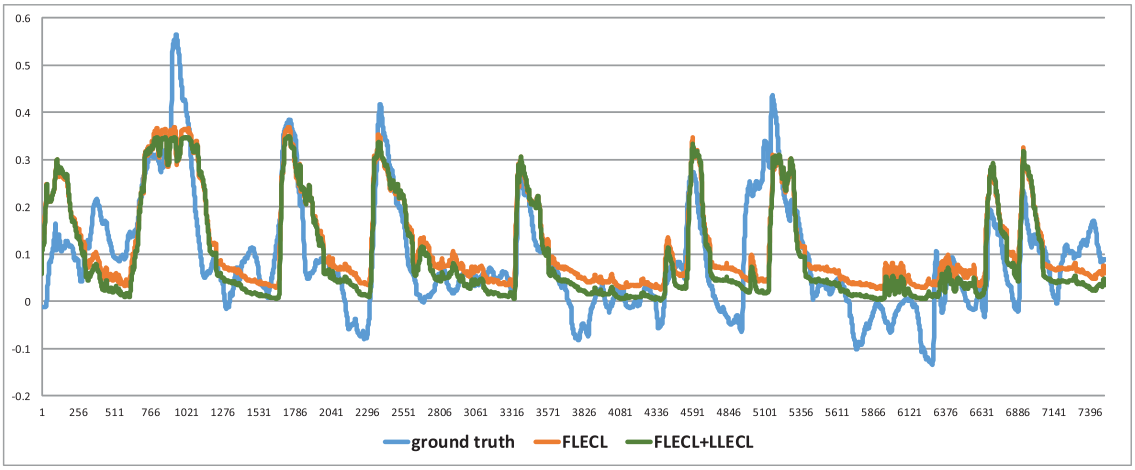


FIGURE 5. **Valence recognition: With and without LLECL stage.**

sampled sequences, the recognition results are more close to the ground truth values. And the recognition results made by the method trained with the original sequences change very fast, whereas the recognition results made by the method trained with the sampled sequences are much smooth.

Fig. 5 provides the comparison results of the methods with and without LLECL stage. From Fig. 5, we can see that the recognition results made by the methods with and without LLECL stage are similar when valence values are larger than 0.1. However, when the valence values are less than 0.1, the recognition results of the method with LLECL are more close to the ground truth. The reason is that the method with LLECL stage learns sufficiently emotion context information in this situation which leads to the improvement of the recognition accuracy. Combining with the results in Table 1, it can further confirm that the model with feature-level context and label-level context gets better results compared with the model only learning feature level context.

V. CONCLUSION & FUTURE WORK

In this paper, we present a novel hierarchical method to learn feature-level and label-level emotional context for predicting affective dimension values from video sequences. Experiment results demonstrate the promising performance of our model. Recognition performance can be significantly improved with the proposed feature-level emotional context learning stage combined with the high-level feature learning stage. And the recognition is smoother by exploring sampled sequence from the original sequence. With the label-level emotional context learning stage, our model can get a more accurate recognition result. In the future, we plan to explore the context information between different modality (e.g., video, audio, and electrocardiogram) so that our model can learn more context information to improve the recognition performance.

REFERENCES

[1] J.-C. Lin, C.-H. Wu, and W.-L. Wei, "Error weighted semi-coupled hidden Markov model for audio-visual emotion recognition," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 142–156, Feb. 2012.

[2] H. Salih and L. Kulkarni, "Study of video based facial expression and emotions recognition methods," in *Proc. Int. Conf. I-SMAC*, Feb. 2017, pp. 692–696.

[3] H. Zhang and M. Xu, "Modeling temporal information using discrete Fourier transform for recognizing emotions in user-generated videos," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 629–633.

[4] D. Grandjean, D. Sander, and K. R. Scherer, "Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization," *Consciousness Cognition*, vol. 17, no. 2, pp. 484–495, 2008.

[5] S. Marsella and J. Gratch, "Computationally modeling human emotion," *Commun. ACM*, vol. 57, no. 12, pp. 56–67, 2014.

[6] H. Gunes, B. Schuller, M. Pantic, and R. Cowie, "Emotion representation, analysis and synthesis in continuous space: A survey," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. Workshops*, Mar. 2011, pp. 827–834.

[7] A. Graves, C. Mayer, M. Wimmer, J. Schmidhuber, and B. Radig, "Facial expression recognition with recurrent neural networks," in *Proc. Int. Workshop Cognition Tech. Syst.*, Munich, Germany, 2008, pp. 1–6.

[8] H. Meng, N. Bianchi-Berthouze, Y. Deng, J. Cheng, and J. P. Cosmas, "Time-delay neural network for continuous emotional dimension prediction from facial expression sequences," *IEEE Trans. Cybern.*, vol. 46, no. 4, pp. 916–929, Apr. 2015.

[9] L. He, D. Jiang, L. Yang, E. Pei, P. Wu, and H. Sahli, "Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks," in *Proc. 5th Int. Workshop Audio/Vis. Emotion Challenge (AVEC)*, Oct. 2015, pp. 73–80.

[10] L. Zhang, D. Tjondronegoro, and V. Chandran, "Representation of facial expression categories in continuous arousal–valence space: Feature and correlation," *Image Vis. Comput.*, vol. 32, no. 12, pp. 1067–1079, 2014.

[11] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Trans. Affective Comput.*, vol. 2, no. 2, pp. 92–105, Apr./Jun. 2011.

[12] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 6, pp. 1113–1133, Jun. 2015.

[13] R. J. Harris, A. W. Young, and T. J. Andrews, "Morphing between expressions dissociates continuous from categorical representations of facial expression in the human brain," *Proc. Nat. Acad. Sci. USA*, vol. 109, no. 51, pp. 21164–21169, 2012.

[14] H. Gunes and M. Pantic, "Dimensional emotion prediction from spontaneous head gestures for interaction with sensitive artificial listeners," in *Proc. Int. Conf. Intell. Virtual Agents*, 2010, pp. 371–377.

[15] L. Shao, X. Zhen, D. Tao, and X. Li, "Spatio-temporal laplacian pyramid coding for action recognition," *IEEE Trans. Cybern.*, vol. 44, no. 6, pp. 817–827, Jun. 2014.

[16] I. Cohen, N. Sebe, A. Garg, L. S. Chen, and T. S. Huang, "Facial expression recognition from video sequences: Temporal and static modeling," *Comput. Vis. Image Understand.*, vol. 91, nos. 1–2, pp. 160–187, 2003.

[17] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden Markov models," *Speech Commun.*, vol. 41, no. 4, pp. 603–623, 2003.

[18] Z. Zeng *et al.*, "Audio-visual affect recognition through multi-stream fused HMM for HCI," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2005, pp. 967–972.

[19] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[20] F. Ringeval *et al.*, "AV+EC 2015: The first affect recognition challenge bridging across audio, video, and physiological data," in *Proc. 5th Int. Workshop Audio/Vis. Emotion Challenge (AVEC)*, Oct. 2015, pp. 3–8.

[21] L. Chao, J. Tao, M. Yang, Y. Li, and Z. Q. Wen, "Long short term memory recurrent neural network based multimodal dimensional emotion recognition," in *Proc. 5th Int. Workshop Audio/Vis. Emotion Challenge (AVEC)*, Oct. 2015, pp. 65–72.

[22] S. Chen and Q. Jin, "Multi-modal dimensional emotion recognition using recurrent neural networks," in *Proc. Int. Workshop Audio/Vis. Emotion Challenge*, 2015, pp. 49–56.

[23] M. Wöllmer *et al.*, "Abandoning emotion classes—Towards continuous emotion recognitionwith modelling of long-range dependencies," in *Proc. INTERSPEECH*, 2008, pp. 597–600.

[24] Z. Huang *et al.*, "An investigation of annotation delay compensation and output-associative fusion for multimodal continuous emotion prediction," in *Proc. 5th Int. Workshop Audio/Vis. Emotion Challenge (AVEC)*, Oct. 2015, pp. 41–48.

[25] H. Meng and N. Bianchi-Berthouze, "Affective state level recognition in naturalistic facial and vocal expressions," *IEEE Trans. Cybern.*, vol. 44, no. 3, pp. 315–328, Apr. 2014.

[26] M. A. Nicolaou, H. Gunes, and M. Pantic, "A multi-layer hybrid framework for dimensional emotion classification," in *Proc. 19th ACM Int. Conf. Multimedia*, 2011, pp. 933–936.

[27] H. Y. Meng and N. Bianchi-Berthouze, "Naturalistic affective expression classification by a multi-stage approach based on hidden Markov models," in *Affective Computing and Intelligent Interaction* (Lecture Notes in Computer Science). Springer, 2011, pp. 378–387.

[28] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Apr. 2013, pp. 1–8.

[29] P. Khorrami, T. Le Paine, K. Brady, C. Dagli, and T. S. Huang, "How deep neural networks can improve emotion recognition on video data," in *Proc. IEEE Int. Conf. Image Process.*, Phoenix, AZ, USA, Sep. 2016, pp. 619–623.

[30] F. Bastien *et al.*, "Theano: New features and speed improvements," in *Proc. Deep Learn. Unsupervised Feature Learn. NIPS Workshop*, 2012.

QIRONG MAO received the M.S. and Ph.D. degrees in computer application technology from Jiangsu University, Zhenjiang, China, in 2002 and 2009, respectively, where she is currently a Professor with the School of Computer Science and Communication Engineering. She has published over 40 technical articles, some of them in premium journals and conferences, such as the IEEE Transactions on Multimedia and ACM Multimedia. Her research interests include affective computing, pattern recognition, and multimedia analysis. Her research is supported by the National Science Foundation of China (NSFC), Jiangsu, and the Education Department of Jiangsu Province.

QING ZHU received the B.E. degree in computer science and technology from Hebei Finance University, Baoding, China, in 2017. She is currently pursuing the M.E. degree with the School of Computer Science and Communication Engineering, Jiangsu University. Her research interests include affective computing, pattern recognition, and deep learning.

QIYU RAO received the B.S. degree in automation (numerical control technology) from the Nanjing Institute of Technology, Nanjing, China, in 2014, and the M.S. degree in computer science and technology from the School of Computer Science and Communication Engineering, Jiangsu University, China, in 2017. His research interests include affect computing and deep learning.

HONGJIE JIA received the Ph.D. degree in computer application technology from the China University of Mining and Technology, Xuzhou, China, in 2017. He is currently a Lecturer with the School of Computer Science and Communication Engineering, Jiangsu University. His research interests include clustering, data mining, and machine learning.

SIDIAN LUO is currently pursuing the B.E. degree with the School of Computer Science and Communication Engineering, Jiangsu University. His research interests include pattern recognition and deep learning.