Emotion Recognition from Multi-Channel EEG through Parallel Convolutional Recurrent Neural Network

Yilong Yang
Software School
Xiamen University
Xiamen, China
yilongyang@stu.xmu.edu.cn

Qingfeng Wu*
Software School
Xiamen University
Xiamen, China
qfwu@xmu.edu.cn

Ming Qiu
Software School
Xiamen University
Xiamen, China
mingqiu@xmu.edu.cn

Yingdong Wang
Software School
Xiamen University
Xiamen, China
yingdongwang@stu.xmu.edu.cn

Xiaowei Chen
Software School
Xiamen University
Xiamen, China
wdenxw@stu.xmu.edu.cn

***Abstract*—As a challenging pattern recognition task, automatic real-time emotion recognition based on multi-channel EEG signals is becoming an important computer-aided method for emotion disorder diagnose in neurology and psychiatry. Traditional machine learning approaches require to design and extract various features from single or multiple channels based on comprehensive domain knowledge. Consequently, these approaches may be an obstacle for non-domain experts. On the contrast, deep learning approaches have been used successfully in many recent literatures to learn features and classify different types of data. In this paper, baseline signals are considered and a simple but effective pre-processing method has been proposed to improve the recognition accuracy. Meanwhile, a hybrid neural network which combines 'Convolutional Neural Network (CNN)' and 'Recurrent Neural Network (RNN)' has been applied to classify human emotion states by effectively learning compositional spatial-temporal representation of raw EEG streams. The CNN module is used to mine the inter-channel correlation among physically adjacent EEG signals by converting the chain-like EEG sequence into 2D-like frame sequence. The LSTM module is adopted to mine contextual information. Experiments are carried out in a segment-level emotion identification task, on the DEAP benchmarking dataset. Our experimental results indicate that the proposed pre-processing method can increase emotion recognition accuracy by 32% approximately and the model achieves a high performance with a mean accuracy of 90.80% and 91.03% on valence and arousal classification task respectively.**

Keywords*—EEG, emotion recognition, deep learning, CNN, RNN

I. INTRODUCTION

Emotion plays an important role in human daily life, it reflects human feelings of things. The mental health status even influences human interpersonal interaction and decision making. In medical fields of psychiatry and neurology, the detected emotional states of patients can be adopted as an indicator of the certain functional emotional disorders, such as posttraumatic stress disorders and major depression. Most recently, researchers analyzed the emotional characteristics of a smartphone overuse group and a healthy group through EEG signals [1].

Human emotions can be detected by facial expressions [2], speech [3], eye blinking [4] and physiological signals [5]. However, the first three approaches are susceptible to subjective influences of the participants, that is, participants can deliberately disguise their emotions. While the physiological signals such as electroencephalograms (EEG), electrooculography (EOG), blood volume pressure (BVP) are produced spontaneously by human body. Consequently, the physiological signals is more objective and reliable in capturing human real emotional states. Of all of these physiological signals, the EEG signal comes directly from human brain, which means changes in EEG signals can directly reflect changes in human emotional states. For this reason, researchers intend to study human emotion through EEG signals.

There already exits extensive studies using machine learning to identify emotion states with EEG signals. Traditional machine learning based methods have shown effective in classifying emotional states, while the shortage of these kinds of approaches is that the researchers must devote numerous efforts to find and design various emotion-related features from origin noisy signals. And the computation of these features is time consuming. A variety of feature extraction methods are proposed in recent years [6]. While the most commonly used methods are Fourier Transform (FT), Power Spectral Density (PSD) and Wavelet Transform (WT) [7]. Chen and Zhang compared two different feature extraction approaches and four different machine learning classifiers and found that nonlinear dynamic features lead to higher accuracy [8]. Yin et al. proposed

* Corresponding author

978-1-5090-6014-6/18/$31.00 ©2018 IEEE

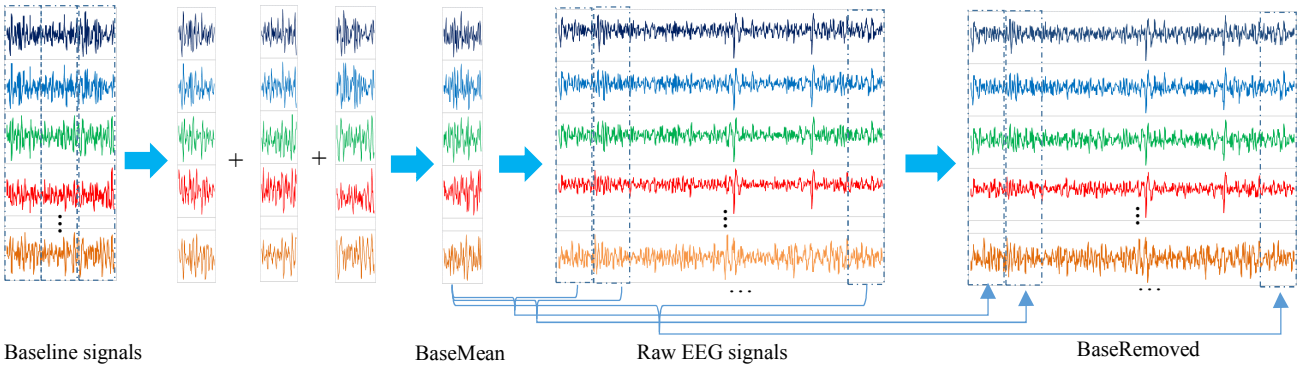

Fig. 1. Flowchart of pre-processing.

a transfer recursive feature elimination (T-RFE) approach that selects features of EEG to determine the optimal feature subset regarding a cross subject emotion classification issue [9].

In recent years, deep learning has drawn wide attention due to its great success in the visual field [10]. Some deep learning based approaches have also achieved competitive accuracy in EEG-based recognition task. In [11], CNN was used to extract time, frequency and location information features of EEG and Stacked Auto Encoders (SAE) was employed to improve the classification accuracy. Zhang et al. proposed both cascade and parallel convolutional recurrent neural network for EEG based movement intention recognition and achieved a perfect performance [12]. Li et al. proposed a pre-processing method that transforms the multi-channel EEG data into 2D frame representation and integrates CNN and RNN to recognition emotion states in the trial-level [13]. Li et al. extracted Power Spectral Density from different EEG channels and mapped it to two-dimensional plane to construct the EEG Multidimensional Feature Image (EEG MFI), then CNN was adopted to learn temporary image patterns from EEG MFI sequences, while the LSTM RNN was used to classify human emotions [14]. Tang et al. used Bimodal Deep Denoising Auto Encoder and Bimodal-LSTM to classify emotion states and achieved the state-of-the-art performance [15].

However, most of CNN based approaches still rely on complex pre-processing and hand-engineered features to a great extent, such as converting raw EEG signals into images [[11],[13],[14]], which may underutilize the ability of deep learning (the features and shared representation can be learned automatically). In addition, Most of EEG-based emotion recognition researches directly employ the EEG signals without taking into account the role of the baseline (EEG signals without stimulation). Hence, to address the issues mentioned above, a simple and computational cheap pre-processing method has been proposed to take the baseline signals into account and finally transform the raw 1D chain-like EEG signals into 2D frame-like sequences. While mapping 1D vectors into 2D frame, the rule is: signals come from physically adjacent channels are still adjacent in the frame, so the spatial information can be retained after converting. Next, a hybrid deep learning structure that integrates the Convolutional Neural Network and Recurrent Neural Network is adopted to conduct emotion recognition tasks in one single framework. Specially, CNN is used to extract spatial features from data frames. RNN is used to extract temporal features from EEG sequence. After the processing of CNN and RNN, a feature fusion method is applied to fuse the spatial features and temporal features. We have evaluated our method on the DEAP [16] dataset and achieved the state-of-the-art performance.

The rest of this paper is organized as follows: A detailed description of the proposed pre-processing method and the hybrid deep learning structure is presented in section 2. Datasets and experiments as well as their results are presented and discussed in section 3. Section 4 highlights the main conclusions of our research.

II. METHODS

A. Pre-processing

In order to improve the recognition accuracy, we use pre-trial data to measure the differences between baseline signals and signals which are recorded while participants are under stimulation. First, we take out pre-trial signals from all C channels and cut it into N segments with a same length L. After the first step, we can get N $(C \times L)$ matrixes. Second, we do the element-wise addition for all of these matrixes and calculate the mean value. This step can be formulated as:

$$\text{BaseMean} = \frac{\sum_1^N mat_i}{N} \tag{1}$$

Here $\text{mat}_i \in \mathbb{R}^{C \times L}$ denotes the ith matrix. After these two steps, we obtain a $C \times L$ matrix named BaseMean, which is used to represent subjects' basic emotional state without any stimulation. Third, we segment the raw EEG signals into M $(C \times L)$ matrixes named RawEEG and then minus the BaseMean for each matrix. The data we use to represent the difference between experiment signals and baseline signals is named as BaseRemoved, it is created as follows:

$$\text{BaseRemoved}_j = \text{RawEEG}_j - \text{BaseMean} \tag{2}$$

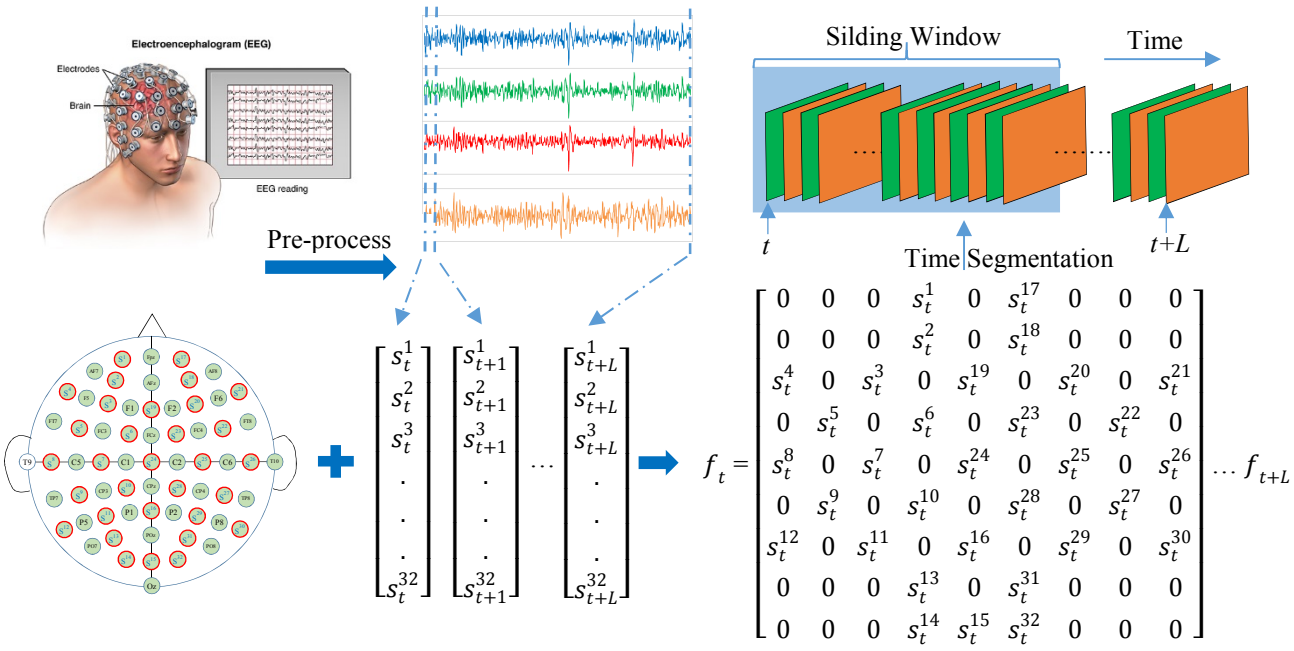

Fig. 2. Converting 1D EEG signals to 2D EEG frames.

The last step is concatenating all of these BaseRemoved matrixes into a big matrix of which size is the same as raw EEG signals. The flowchart of pre-processing is shown in Fig. 1.

B. Converting 1D EEG signals into 2D EEG frames

The overall EEG data acquisition and transformation flowchart is shown in Fig. 2. The EEG based BCI system uses a wearable headset with multiple electrodes to capture EEG signals. The International 10-20 System is an internationally recognized method of describing and applying the location of scalp electrode and the underlying area of the cerebral cortex. The "10" and "20" refer to the fact that the actual distance between the adjacent electrodes are either 10% or 20% of the total front front-back or right-left distance of the skull.

The data from the EEG signal acquisition system at time index t is a 1D data vector $\boldsymbol{v}_t = [s_t^1, s_t^2, s_t^3, \ldots, s_t^n]^T$. Where s_t^i is a pre-processed data of the ith electrode channel and the acquisition system totally contains n channels. With the DEAP dataset, n equals 32. For the observation period $[t, t + L]$, there are $(L + 1)$ 1D data vectors, each of which contains n elements corresponding to n electrodes of the acquisition headset. We can see that the lower left corner of Fig. 2 is the plan view of the International 10-20 System, where the EEG electrodes circled in red are the test points used in the DEAP dataset. In this study, we generalized the International 10-20 System with test electrodes used in the DEAP dataset to form a matrix $(h \times w)$, where h is the maximum point number of the vertical test points and w is the maximum point number of the horizontal test points. With the DEAP dataset, h equals w equals 9. From the EEG electrode map, each electrode is physically neighboring multiple electrodes which records the EEG signals in a certain area of the brain, while the elements of the chain-like 1D EEG data vector are restricted to two neighbors. For the purpose of maintaining spatial information among multiple adjacent channels, we convert the 1D EEG data vectors to 2D EEG frames according to the electrode distribution map. The corresponding 2D data frame $\boldsymbol{f}_t$ of the 1D data vector $\boldsymbol{v}_t$ at time index t is denoted as follows:

$$\boldsymbol{f}_t = \begin{bmatrix} 0 & 0 & 0 & s_t^1 & 0 & s_t^{17} & 0 & 0 & 0 \\ 0 & 0 & 0 & s_t^2 & 0 & s_t^{18} & 0 & 0 & 0 \\ s_t^4 & 0 & s_t^3 & 0 & s_t^{19} & 0 & s_t^{20} & 0 & s_t^{21} \\ 0 & s_t^5 & 0 & s_t^6 & 0 & s_t^{23} & 0 & s_t^{22} & 0 \\ s_t^8 & 0 & s_t^7 & 0 & s_t^{24} & 0 & s_t^{25} & 0 & s_t^{26} \\ 0 & s_t^9 & 0 & s_t^{10} & 0 & s_t^{28} & 0 & s_t^{27} & 0 \\ s_t^{12} & 0 & s_t^{11} & 0 & s_t^{16} & 0 & s_t^{29} & 0 & s_t^{30} \\ 0 & 0 & 0 & s_t^{13} & 0 & s_t^{31} & 0 & 0 & 0 \\ 0 & 0 & 0 & s_t^{14} & s_t^{15} & s_t^{32} & 0 & 0 & 0 \end{bmatrix} \quad (3)$$

Zero is used to represent the signals from the channels that are unused in DEAP dataset, which has no effects on neural network. By this transformation, the pre-processed 1D data vector sequences $[\boldsymbol{v}_t, \boldsymbol{v}_{t+1}, \ldots, \boldsymbol{v}_{t+L}]$ is converted to 2D data frame sequences $[\boldsymbol{f}_t, \boldsymbol{f}_{t+1}, \ldots, \boldsymbol{f}_{t+L}]$. For the observation duration $[t, t + L]$, the quantity of 2D data frames is still $(L + 1)$. After transformation, each data frame is normalized across the non-zero elements using Z-score normalization by the following equation:

$$z = \frac{x - \mu}{\sigma} \quad (4)$$

Here x denotes a non-zero element from a certain position of the frame, μ denotes the mean of all non-zero elements and σ denotes the stand deviation of these elements. Finally, we apply the sliding window approach to segment the streaming 2D frames to individual frame-group as shown in the last step of Fig.

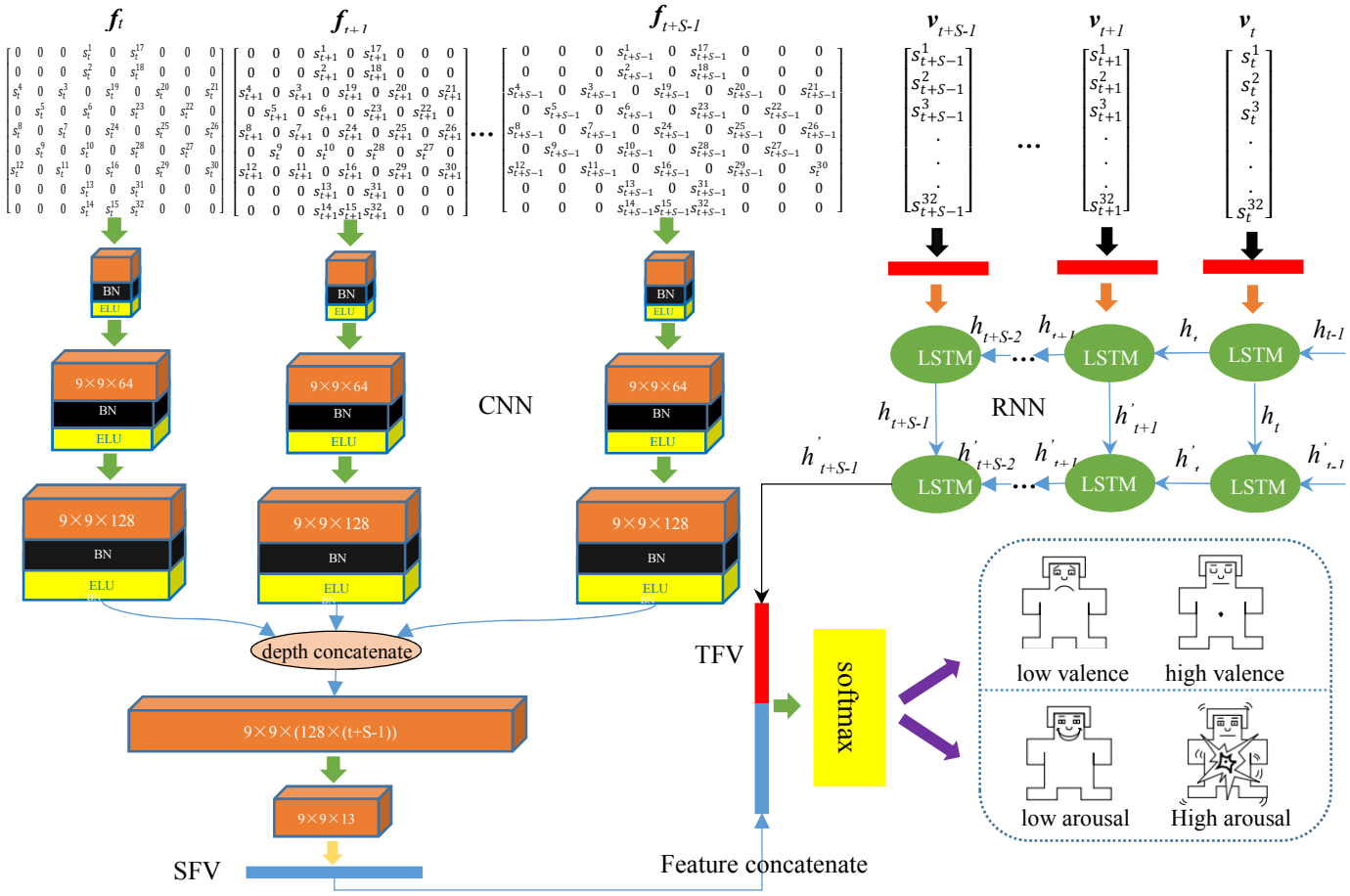

Fig. 3. Parallel Convolutional Recurrent Neural Network.

2. Each frame-group is a sequence of 2D frames with a fixed length without any overlap between consecutive neighbors. The data frames segment $\mathbf{S}_j$ is created as follows:

$$\mathbf{S}_j = [\boldsymbol{f}_t, \boldsymbol{f}_{t+1}, \ldots, \boldsymbol{f}_{t+S-1}] \tag{5}$$

Where S denotes the window size and subscript j is used to identify the different segments during the observation period. The goal of this paper is to develop an effective model to recognize a set of human emotions $\mathbf{E} = [e_1, e_2, \ldots, e_n]^T$ from each windowed data frames segment $\mathbf{S}_j$.

C. Parallel Convolutional Recurrent Neural Network

Besides the pre-processing method mentioned above, we also adopt a hybrid deep learning model to classify emotion states, named "Parallel Convolutional Recurrent Neural Network". The model is a composition of two kinds of deep learning structures. It combines the powerful ability of CNN and RNN in extracting spatial and temporal features respectively. The CNN unit works for mining cross-channel correlation and extracting features from 2D frames. The refined RNN structure named "Long Short-Term Memory (LSTM)" models the context information for streaming 1D data vectors. Following these two units, a feature fusion method is used to fuse the extracted features at last for final emotion recognition. The structure of the parallel convolutional recurrent neural network is depicted in Fig. 3.

For CNN part, there are three continuous 2D convolutional layers with a same kernel size of 4×4 for spatial feature extraction. While the 3×3 kernel is widely used in computer vision field, we choose 4×4 filter as the signals in 2D EEG frame is sparse. So 4 by 4 filter can mine the correlation among more channels than 3 by 3 kernel. In each convolutional layers, we use zero-padding to prevent missing information at the edge of input data frame. Then we start the first convolutional layer with 32 feature maps and double the feature maps in each of the following convolutional layers. Hence, there are 64 and 128 feature maps in the second and the third convolutional layers respectively. Although a convolutional layer is often followed by a pooling layer in classical CNN architectures, it is not necessary in our model. The pooling layer is usually added for reducing data dimensional at the cost of missing some information. While in this EEG recognition task, the size of data frame is much smaller than that used in computer vision field. Thus, in order to keep all information, we do not use pooling operation in our model. Following each convolution operation, a batch normalization (BN) operation is applied to accelerate the model training. After these three convolutional layers, for the purpose of fusing spatial feature vectors and temporal feature

vector which is extracted by RNN, a depth concatenate operation is applied to combine the spatial feature maps into a large cube. And then we use 13 1×1 convolution kernels to shrink the cube into $13\times9\times9$, and further flatten it into a spatial feature vector $\mathbf{SFV} \in \mathbb{R}^{1053}$. The purpose of 1 by 1 convolution operation is to fuse feature maps at different times and play a role in dimensionality reduction.

The input of CNN part is a pre-processed segment of 2D data frame, while due to the RNN part being responsible for the temporal feature extraction, 1D data vectors are not converted to 2D frame sequences. The jth input segment to the CNN is denoted as $S_j = [\boldsymbol{f}_t, \boldsymbol{f}_{t+1}, \ldots, \boldsymbol{f}_{t+S-1}] \in \mathbb{R}^{S\times h\times w}$, where there are S data frames and each of them denoted as $\boldsymbol{f}_k$ ($k = t, t+1, \ldots, t+S-1$). Each segment is fed into 2D-CNN and resolved to a Spatial Feature Vector **SFV**:

$$\mathbf{SFV}_j = Conv2D(S_j), \mathbf{SFV}_j \in \mathbb{R}^{1053} \tag{6}$$

For RNN part, we adopt Long Short-Term Memory unit to construct two stacked RNN layers. There are S LSTM units in each RNN layer due to the same window size, and the input of the second RNN layer is the output time sequence of the first RNN layer. The hidden state of the LSTM unit in the first layer at current time step t is denoted as h_t, and the h_{t-1} is the hidden state of the previous time step $t-1$. The information from the previous time step is conveyed to the current time step and influences the final output. We use the hidden state of the LSTM unit as its output. Therefore, the input sequence of the second LSTM layer is the hidden state sequence of the first LSTM layer $[\mathbf{h}_t, \mathbf{h}_{t+1}, \ldots, \mathbf{h}_{t+S-1}]$. Since we focus on segment-level emotion recognition rather than time step level, only the output of the last time step, $\mathbf{h}'_{t+S-1}$, is fed into the next fully connected layer. While due to the RNN part being used to extract temporal features, the 1D EEG data vectors are not transformed to 2D frames. The jth inputted windowed segment to the RNN part is:

$$\mathbf{R}_j = [\boldsymbol{v}_t, \boldsymbol{v}_{t+1}, \ldots, \boldsymbol{v}_{t+S-1}] \tag{7}$$

Where $\boldsymbol{v}_t$ is the vector at time step t, and S denotes the window size. The hidden state of the last time step in one segment is:

$$\mathbf{h}'_{t+S-1} = LSTM(\mathbf{R}_j), \mathbf{h}'_{t+S-1} \in \mathbb{R}^d \tag{8}$$

Where d is the hidden state size of the LSTM unit. A fully connected layer is applied both before and after LSTM layers to enhance the temporal information representation capability. Therefore, the final Temporal Feature Vector ($\mathbf{TFV}_j$) of segment $\mathbf{R}_j$ is denoted as:

$$\mathbf{TFV}_j = FC(\mathbf{h}'_{t+S-1}), \mathbf{TFV}_j \in \mathbb{R}^l \tag{9}$$

Where l is the size of the final fully connected layer. Finally, the concurrently extracted spatial and temporal features are concatenated to a joint spatial-temporal feature vector and a softmax layer receives it as an input to predict human emotion states:

TABLE I. DATA FORMAT

Array name	*array shape*	*Array contents*
data	40 × 40 × 8064	video/trial × channel × data
labels	40×4	video/trial × label (valence, arousal, dominance, liking)

$$\mathbf{P}_j = Softmax([\mathbf{SFV}_j, \mathbf{TFV}_j]), \mathbf{P}_j \in \mathbb{R}^n \tag{10}$$

Where n is the quantity of classes, in our experiment, n is 2. In order to avoid overfitting, we apply dropout operation as a form of regularization after fully connected layers in RNN part. In addition, a L2 regularization term is also added to cost function to improve the generalization ability of the model.

III. EXPERIMENTS

A. The Datasets

In this paper, we use DEAP dataset to validate our proposed approach. The DEAP dataset was first introduced in [16]. EEG signals and peripheral physiological signals of 32 participants were recorded when they were watching 40 pieces of music videos. The dataset contains 32 channel EEG signals and 8 channel peripheral physiological signals. Here the EEG signals are used for emotional recognition and the peripheral physiological signals are abnegated. During the experiment, the EEG signals were sampled at 512 Hz and then down-sampled to 128 Hz. EOG artefacts were removed. A bandpass frequency filter from 4.0-45.0 Hz was applied. The preprocessed EEG data contains 60s trial data and 3s baseline data. The emotional music videos include 40 one-minute long clips and participants were asked to rate the levels of arousal, valence, liking and dominance for each video. Each subject file contains two arrays, the data format of file is illustrated in Table I. In order to compare the performance of our proposed method with previous result in [9], [15], [17], [18], [19], we choose 5 as threshold to divide the trials into two classes according to the rated levels of arousal and valence. Then the task can be treated as two binary classification problems, namely high or low arousal and valence.

B. Model Implementation

An appropriate time window length is critical for classification performance on streaming data. Wang et al. have found that 1 second long time window is the most suitable window length for emotion recognition [20]. Hence, 1 s is chosen as the time window length in this paper. Since the signals were down-sampled to 128 Hz, the pre-trial baseline signals are segmented to 3 32×128 matrixes to calculate the BaseMean matrix. Then the pre-process operation is applied to trial signals, as shown in Fig. 1. After the pre-processing stage, in each trial, we get 32 channels' pre-processed EEG signals and use a sliding window to divide it into 60 segments with 1 s length, which

TABLE II. PERFORMANCE CONPARISON BETWEEN USING BASELINE AND WITHOUT BASELINE

Recognition Accuracy (%) Comparison for Each Subject on "Valence"(pre-process with baseline and without pre-process)											
Sub	***without***	***with***	***Sub***	***without***	***with***	***Sub***	***without***	***with***	***Sub***	***without***	***with***
1	50.75	**92.93**	9	54.25	**88.70**	17	48.00	**84.57**	25	53.50	**90.85**
2	57.50	**85.07**	10	65.25	**90.05**	18	62.25	**90.32**	26	62.25	**86.25**
3	47.00	**94.80**	11	51.50	**83.67**	19	41.75	**90.65**	27	70.75	**94.25**
4	54.50	**85.42**	12	50.75	**92.90**	20	54.75	**94.60**	28	66.25	**89.87**
5	55.25	**88.90**	13	55.00	**93.90**	21	58.00	**93.60**	29	52.50	**93.20**
6	70.75	**90.92**	14	60.25	**90.10**	22	56.25	**91.23**	30	73.75	**91.62**
7	58.75	**92.87**	15	55.25	**92.15**	23	65.00	**94.70**	31	55.50	**90.35**
8	53.00	**92.32**	16	55.75	**90.85**	24	54.00	**94.18**	32	55.75	**89.63**
Recognition Accuracy (%) Comparison for Each Subject on "Arousal"(pre-process with baseline and without pre-process)											
Sub	***without***	***with***	***Sub***	***without***	***with***	***Sub***	***without***	***with***	***Sub***	***without***	***with***
1	58.75	**93.00**	9	53.50	**88.35**	17	53.00	**85.75**	25	65.25	**91.78**
2	58.75	**86.68**	10	53.50	**89.85**	18	58.00	**91.92**	26	53.00	**86.50**
3	76.75	**95.45**	11	60.25	**85.03**	19	66.75	**90.95**	27	67.25	**94.37**
4	53.25	**84.78**	12	73.75	**93.43**	20	73.75	**94.48**	28	58.25	**90.47**
5	55.00	**88.40**	13	85.25	**94.53**	21	77.50	**91.63**	29	59.75	**94.05**
6	63.25	**90.10**	14	56.00	**90.37**	22	49.00	**90.50**	30	61.25	**93.40**
7	61.00	**90.68**	15	54.50	**93.38**	23	76.25	**94.70**	31	48.00	**89.30**
8	61.00	**92.55**	16	47.00	**91.55**	24	73.75	**94.42**	32	64.75	**90.57**
Average Recognition Accuracy Results across Subjects on "Valence" and "Arousal" (mean± std. dev.)											
				Valence			***Arousal***				
without pre-process				57.05 ± 7.10			61.78 ± 9.61				
pre-process with baseline				**90.80 ± 3.08**			**91.03 ± 2.99**				

means the window size S is 128. After segmenting, we get a total of 2400 samples (40 trials×60 segments) for each subject. Then the 2D data frames are transformed with the size of 9×9 as shown in Fig. 2. We use10-fold cross-validation to evaluate the performance of our approach. The average performance of the 10-fold validation processes is taken as the experiment's final results.

The model was implemented with Tensorflow framework and trained on an NVIDIA TITAN XP pascal GPU. The Adam optimizer is adopted to minimize the cross-entropy loss function. The keep probability of dropout operation is 0.5. The penalty strength of L2 is 0.5. The hidden sates of the LSTM cell d is 32. All fully connected layers have the same size of 1024. The initial learning rate is 10^{-4}, while the training accuracy surpasses 80% but less than 85%, the learning rate is set to 5×10^{-5}, while the training accuracy surpasses 85%, it was changed to 5×10^{-6}.

C. Results

In order to validate the effectiveness of our proposed pre-processing method, we conduct two experiments. The first one is to use raw EEG signals to perform the recognition task without taking into account the role of the 3-second-long baseline signals. And the other one is to apply the proposed pre-process method before feeding them into model.

As shown in Table II, the proposed pre-processing method can significantly improve the recognition accuracy by nearly 33% and 30% both on valence and arousal recognition task, which indicates the high effectiveness of this approach. The pre-processing method takes the baseline signals as a basic emotional representation state first, then calculates the difference between it and real emotion signals, and uses this difference to represent the emotion state. The experiment result shows this representation is useful. Otherwise, the standard deviation of recognition accuracy is smaller than the experiment that without applying the pre-process.

We also compare our model with five different approaches on DEAP dataset. The mean accuracy of 10 folds cross-validation is used. Li et al. used a deep belief network (DBN) to automatically extract high-level features from raw EEG signals [17]. Atkinson and Campos employed mutual information minimization technique and one-against-one support vector machines (SVMs) as the emotion classifier [18]. Yin et al. developed a transfer recursive feature elimination (T-RFE) to determine a set of the most robust EEG indicators with stable geometrical distribution across a group of training subjects and a specific testing subject [9]. Tang et al. developed a Bimodal-LSTM model to conduct emotion recognition task using both EEG signals and peripheral physiological signals and achieved a state-of-the-art result with a mean accuracy of 83.53% on

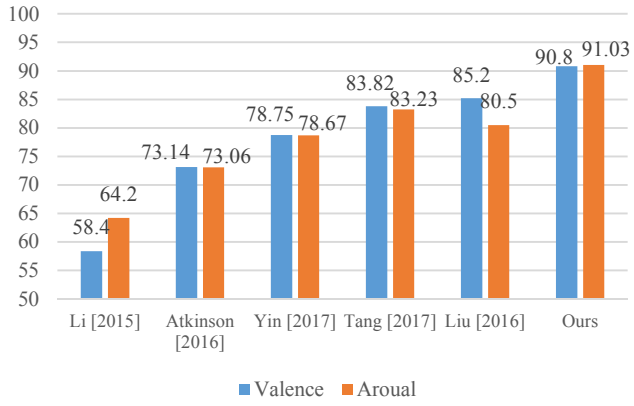

Fig. 4. Performance comparison between relevant approaches.

DEAP dataset [15]. Liu et al. applied a Bimodal Deep AutoEncoder (BDAE) on EEG signals and Eye Signals and achieved mean accuracy of 82.85% when classifying low and high valence and arousal.

The contrastive results on the DEAP dataset are shown in Fig. 4. The comparison shows the effectiveness of our model. The proposed model outperforms the EEG-based only approaches significantly, which is about 30% points higher than Li [2015], 18% points higher than Atkinson [2016] and 12% points higher than Yin [2017]. The performance of our model also surpasses these methods adopted by Liu [2016] and Tang [2017]. While compared to these two approaches, both of them take the eye movement data into account and both Power Spectral Density (PSD) and Differential Entropy (DE) features are extracted from EEG data and eye movement data, which means, we achieve the higher performance but use less data. Otherwise, compared to our pre-processing method, PSD and DE features are time consuming.

IV. CONCLUSION

In this paper, the baseline signals were took into account and a simple, computation cheap pre-processing method has been proposed to improve recognition accuracy. In addition, we applied a hybrid neural network to classify human emotion states by effectively learning compositional spatial-temporal representation of raw EEG streams. At last, the public DEAP dataset was used to evaluate the proposed pre-processing method and neural network model. Experimental results have shown that the pre-processing approach can improve the accuracy by 32% and the model could achieve a high accuracy around 90.80% and 91.03% for valence and arousal classification task, respectively.

ACKNOWLEDGMENT

This work was supported by NSFC (No. 61402387, No. 61402390); the Key Program of Science and Technology of Fujian Province of China (No. 2014H0044); Science and Technology Guiding Project of Fujian Province of China (No.2015H0037, No.2016H0035); Enterprise Technology Innovation Project of Fujian Province; the Education and Research Project of Middle and Young Teacher of Fujian Province of China (No.JA15018); the Overseas Study Scholarship of Fujian Province; Science and Technology Project of Xiamen, China (No. 3502Z20153026); the National Key Research and Development Program of China (No. 2017YFC1703303). The authors would like to thank the researchers (Dalin Zhang et al.) from the School of Computer Science and Engineering, University of New South Wales, who proposed spatial-temporal representation of raw EEG streams to recognize human intention and achieve a state-of-the-art result in motor imagery EEG (MI-EEG), their support is sincerely appreciated.

REFERENCES

[1] Kim, Seul-Kee, and Hang-Bong Kang. "An analysis of smartphone overuse recognition in terms of emotions using brainwaves and deep learning." Neurocomputing (2017).

[2] Anderson, Keith, and Peter W. McOwan. "A real-time automated system for the recognition of human facial expressions." IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 36.1 (2006): 96-105.

[3] Petrushin, Valery. "Emotion in speech: Recognition and application to call centers." Proceedings of Artificial Neural Networks in Engineering. Vol. 710. 1999.

[4] Soleymani, Mohammad, Maja Pantic, and Thierry Pun. "Multimodal emotion recognition in response to videos." IEEE transactions on affective computing 3.2 (2012): 211-223.

[5] Yin, Zhong, et al. "Recognition of emotions using multimodal physiological signals and an ensemble deep learning model." Computer Methods and Programs in Biomedicine 140 (2017): 93-110.

[6] Jenke, Robert, Angelika Peer, and Martin Buss. "Feature extraction and selection for emotion recognition from EEG." IEEE Transactions on Affective Computing 5.3 (2014): 327-339.

[7] Alarcao, Soraia M., and Manuel J. Fonseca. "Emotions recognition using EEG signals: a survey." IEEE Transactions on Affective Computing (2017).

[8] Chen, Peng, and Jianhua Zhang. "Performance Comparison of Machine Learning Algorithms for EEG-Signal-Based Emotion Recognition." International Conference on Artificial Neural Networks. Springer, Cham, 2017.

[9] Yin, Zhong, et al. "Cross-subject EEG feature selection for emotion recognition using transfer recursive feature elimination." Frontiers in neurorobotics 11 (2017).

[10] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." Nature 521.7553 (2015): 436-444.

[11] Tabar, Yousef Rezaei, and Ugur Halici. "A novel deep learning approach for classification of EEG motor imagery signals." Journal of neural engineering 14.1 (2016): 016003.

[12] Zhang, Dalin, et al. "EEG-based Intention Recognition from Spatio-Temporal Representations via Cascade and Parallel Convolutional Recurrent Neural Networks." arXiv preprint arXiv:1708.06578 (2017). In press.

[13] Li, Xiang, et al. "Emotion recognition from multi-channel EEG data through Convolutional Recurrent Neural Network." Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on. IEEE, 2016.

[14] Li, Youjun, et al. "Human Emotion Recognition with Electroencephalographic Multidimensional Features by Hybrid Deep Neural Networks." Applied Sciences 7.10 (2017): 1060.

[15] Tang, Hao, et al. "Multimodal Emotion Recognition Using Deep Neural Networks." International Conference on Neural Information Processing. Springer, Cham, 2017.

[16] Koelstra, Sander, et al. "Deap: A database for emotion analysis; using physiological signals." IEEE Transactions on Affective Computing 3.1 (2012): 18-31.

[17] Li, Xiang, et al. "EEG based emotion identification using unsupervised deep feature learning." (2015).

[18] Atkinson, John, and Daniel Campos. "Improving BCI-based emotion recognition by combining EEG feature selection and kernel classifiers." Expert Systems with Applications 47 (2016): 35-41.

[19] Liu, Wei, Wei-Long Zheng, and Bao-Liang Lu. "Emotion recognition using multimodal deep learning." International Conference on Neural Information Processing. Springer International Publishing, 2016.

[20] Wang, Xiao-Wei, Dan Nie, and Bao-Liang Lu. "Emotional state classification from EEG data using machine learning approach." Neurocomputing 129 (2014): 94-106.