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

Xiang Li[1], Dawei Song[1,2,*], Peng Zhang[1,*], Guangliang Yu[1], Yuexian Hou[1], Bin Hu[3]
[1]Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin University, China
[2]School of Computing and Communications, The Open University, United Kindom
[3]School of Information Science and Engineering, Lanzhou University, China
{xlee, dwsong, pzhang, glyu, yxhou}@tju.edu.cn, bh@lzu.edu.cn

***Abstract*—Automatic emotion recognition based on multi-channel neurophysiological signals, as a challenging pattern recognition task, is becoming an important computer-aided method for emotional disorder diagnoses in neurology and psychiatry. Traditional approaches require designing and extracting a range of features from single or multiple channel signals based on extensive domain knowledge. This may be an obstacle for non-domain experts. Moreover, traditional feature fusion method can not fully utilize correlation information between different channels. In this paper, we propose a preprocessing method that encapsulates the multi-channel neurophysiological signals into grid-like frames through wavelet and scalogram transform. We further design a hybrid deep learning model that combines the '*Convolutional Neural Network (CNN)*' and '*Recurrent Neural Network (RNN)*', for extracting task-related features, mining inter-channel correlation and incorporating contextual information from those frames. Experiments are carried out, in a trial-level emotion recognition task, on the DEAP benchmarking dataset. Our results demonstrate the effectiveness of the proposed methods, with respect to the emotional dimensions of Valence and Arousal.**

***Index Terms*—CNN, EEG, emotion recognition, LSTM, physiological signal.**

I. Introduction

Emotions play an important role in human interpersonal interaction and decision making. Automatic emotion recognition, as a challenging pattern recognition task, has drawn increasing interests from various research fields [15]. A representative field is '*Affective Computing*', which aims to empower computer systems with the ability to automatically recognize, comprehend and respond to human emotions for intelligent human computer interactions. More recently, in medical domains of psychiatry and neurology, the detected emotional states of patients can be adopted as an indicator of certain organic or functional emotional disorders, such as post-traumatic stress disorder and major depression. Traditionally, those symptoms are clinically diagnosed by doctors through interviews with patients based on the DSM-IV criteria [1] or through patients' self-reporting information. Nevertheless, the diagnosis accuracy may be affected by various factors, including the proficiency of the doctors and the cooperation of the patients. Therefore, a computer-aided recognition, combined with traditional clinical methods, would help reach a better diagnosis [23].

Automatically recognizing emotions based on data gathered from various neurophysiological signal acquisition apparatuses, such as multi-channel EEG, is an emerging way for emotion monitoring. Compared with facial expressions or voice based approaches, the neurophysiological signal based approach is more reliable in capturing human's real emotional states. More and more portable physiological signal acquisition apparatuses are available for long-term daily emotion monitoring, making this research direction feasible and meaningful. Indeed, some psychophysiological studies have revealed strong relationships between explicit physiological activities and implicit psychological experiences [4].

Despite of the recent advances, there exist a number of challenges in neurophysiological based emotion recognition, summarized as follows:

- Firstly, numerous efforts have been devoted to finding and designing the various emotion-related features from the weak and noisy signals, such as power spectral density [13], differential entropy [7], asymmetric spatial pattern [10], and high order zero crossing count [16]. Although the effectiveness of some domain-specific, hand-engineered features has been validated, the design of features needs more in-depth study from the perspective of emotion related cognitive research. The computation of those features can be time consuming, especially when extracting features based on theory of chaos and non-linear dynamics (e.g., fractal dimension, correlation dimension and complexity) [21].
- Secondly, capturing the correlation between multiple channel signals is crucial, but typically done through feature-level fusion based approaches [5], multi-channel decision fusion based strategies [11] or deep representation learning based methods [12]. However, the processes of feature extraction and correlation modeling in these methods are handled separately. Recent '*Deep Learning*

*Corresponding author

(DL)' based approaches, although promising, still largely rely on hand-engineered features, which may under-utilize the ability of DL, in which the task-related features and shared representation can be learned automatically.

- Thirdly, traditional machine learning based methods have been shown effective in classifying emotional states rather than modeling the transition and evolution of states. However, the evolving process of emotional states is important for doctors to review and assess a patient's past condition. In addition, the offline machine learning methods are not suitable for incremental learning scenarios, where medical data may be continuously acquired online rather than be provided in advance.
- Finally, most existing work has concentrated on segment-level emotion recognition tasks, where the emotion prediction is conducted for signal segments of 1 second or a little longer. It can not meet the need of long-term emotion monitoring, for which the acquired signals may last for hours or days.

To address the issues mentioned above, we propose a preprocessing method that encapsulates the multi-channel neurophysiological signals into grid-like frames. Each frame represents the wavelet spectral energy information of the multi-channel signals within a specific time window. Further, we propose a hybrid deep learning structure that integrates the Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) for processing the acquired frames and conducting emotion recognition tasks in one single framework. Specifically, the CNN is used for learning task-related features and mining inter-channel correlation from the frames through designed convolutional filters. The RNN is used for modeling the evolution, transition and long term dependencies of the signals in an entire user trial for final emotion prediction. Our method is based on two assumptions: (1) the emotional experience is a reaction to external events and evolves continuously with respect to the change of stimuli, and (2) the neurophysiological signals contain rich contextual and semantic information that is suitable for the RNN to model. We have validated the effectiveness of our method on the DEAP dataset, which is a widely used benchmark for emotion recognition, and also shown that our method has a good potential for realtime prediction.

The rest of this paper is organized as follows. A detailed description of the proposed preprocessing method and the DL structure is presented in Section 2. Section 3 decribes the experimental setup and reports the experimental results. Finally, we conclude the paper in Section 4.

II. The Methods for Multi-Channel EEG Based Emotion Recognition

Our proposed methodology addresses two research problems. The first one is how to preprocess and represent the multi-channel signals before adopting some modeling methods. The second one is how to model and recognize emotions based on the preprocessed data. Correspondingly, we propose a 2D frame based representation method and a hybrid deep learning framework, respectively. The brief process of our framework in dealing with multi-channel neurophysiological signal based emotion recognition is illustrated in Figure 1.

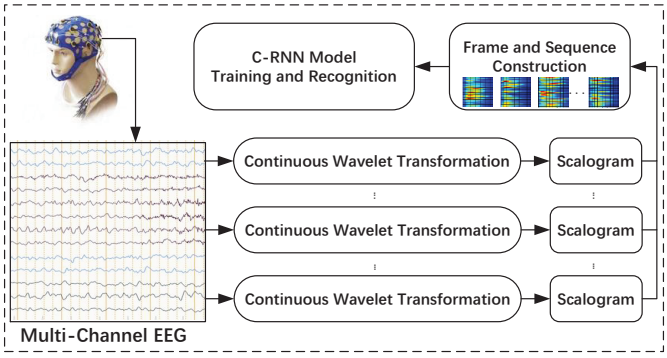

Fig. 1: The outlines and process of our framework designed for multi-channel EEG based emotion recognition.

A. Data Preprocessing and Frame Construction Method

In this work, we construct two-dimensional data structure which is called '*frame*' by us, denoted as M_t, each represents the wavelet energy distribution in dimensions of '*channel*' and '*scale*' (a terminology in signal processing, each scale represents a specific coarseness of the signal) in the time window t. The advantage of the frame structure is that it encapsulates and integrates information of all channels' signal into a direct-viewing form. Therefore, these multi-channel signals can be further processed as a whole, and the inner relationship of the multiple channels can be mined, especially suitable for the multi-channel EEG signal based data mining tasks, where each channel's signal is the mixture of electrical activity arising from various cerebral regions distributed in the brain. Furthermore, the sequence of frames $< M_1, M_2, ..., M_n >$ reflects the dynamic activity changing in different cortical areas during an emotional experience.

Before constructing the frames, we firstly need to conduct '*Continuous Wavelet Transform (CWT)*' for each-channel signal, and we then need further transform the output from CWT into scalograms, as detailed below.

1) Wavelet Based Sparse Representation: The representation and approximation of a signal is the key issue in signal processing and related pattern recognition tasks. The '*Sparse Representation (SR)*' plays important roles in fields including signal processing, machine learning, computer vision, etc. It helps learn a compact structure and get high-level implicit semantic information from raw signals [18]. '*Wavelet Transform (WT)*' is typically regarded as a kind of SR. It is excellent in denoting local transitory characteristics in both frequency and time domain. Therefore, it is quite suitable to be applied to process non-stationary neurophysiological signals, such as EEG [22]. Compared with the wavelet transform, traditional '*Windowed Fourier Transform*' based time-frequency analysis (e.g., the *'Short Time Fourier Transform (STFT)'*) is only suitable for processing stationary signal and it is incapable of getting a high joint frequency-time resolution according to

the '*Heisenberg's Uncertainty Principle*'. Besides the intrinsic advantages, why we adopt WT rather than directly using raw signals to construct frames is because the size of a frame will be large, especially when the raw signals' sampling rate is high. Therefore, we take the wavelet based SR into consideration.

Compared with traditional SR methods, such as '*Sparse Coding*'. The dictionary of wavelet analysis is not acquired through learning, but is predetermined by a mother wavelet ψ. After scaling s and translation u of the mother wavelet a group of wavelet basis functions $\psi_{s,u}$ can be acquired, as Formula 1.

$$\psi_{s,u}(t) := \frac{1}{\sqrt{s}}\psi(\frac{t-u}{s}), u \in \mathbb{R}, s > 0. \quad (1)$$

The raw channel signal is decomposed according to these basis functions. Therefore, the effects of representation is largely affected by what kind of mother wavelet is selected. You have to deliberately choose from various kinds of wavelet families, such as Haar wavelet, Daubechies wavelet, Symlet wavelet, Coif Wavelet, Bior Wavelet, etc [8]. In this work we choose the Db-4 wavelet to do CWT for each channel signal, which is formulated as Formula 2, where $f(t)$ is the original EEG signal.

$$Wf(s,u) := \int_{-\infty}^{+\infty} f(t)\bar{\psi}_{s,u}(t)dt. \quad (2)$$

After the CWT, each one-dimensional channel signal is transformed into a wavelet coefficients based time-scale representation, as shown in Figure 2. The notion of scale is introduced as an alternative of frequency. Each scale corresponds to a scaled version of the mother wavelet, where the low scale is corresponding to high-frequency component of the signal and the high scale is corresponding to low-frequency component. The number of scales we specify is determined according to the properties of the raw signal, such as the sampling rate and the the cutoff frequency. For example, if the sampling rate is 128Hz, then we can obtain at most 64 frequency components from the raw signal (according to the '*Nyquist's Sampling Theorem*'). Each element in the time-scale representation is the calculated wavelet coefficient corresponding to a specific scale of mother wavelet. The wavelet coefficients can also be used to reconstruct the original signal, so the wavelet coefficients here are regarded as the signal's alternative representation.

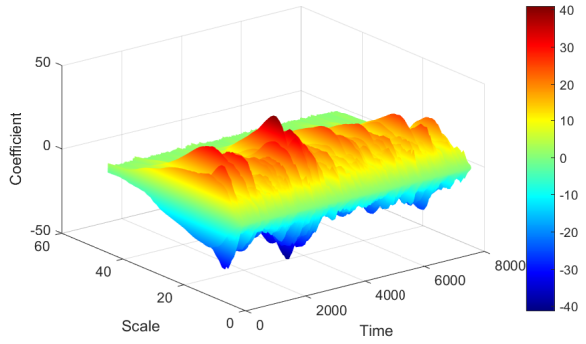

Fig. 2: The time-scale representation of the wavelet coefficients obtained after CWT for a channel signal. The larger the absolute value of the wavelet coefficient, the greater the proportion of the corresponding component in this channel signal.

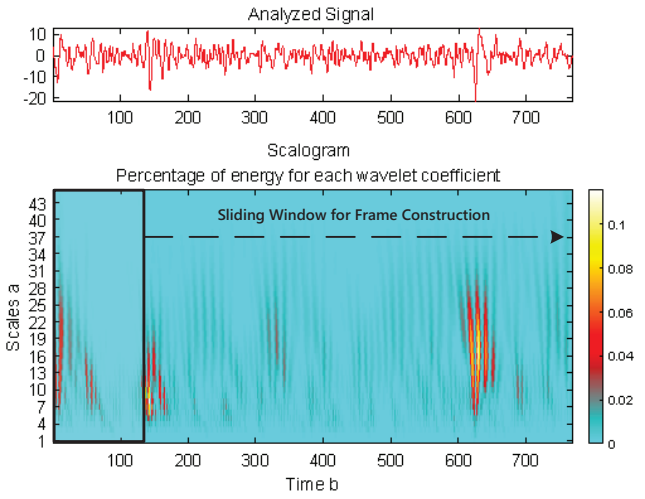

Fig. 3: We further transform one channel signal into a scalogram based on the obtained wavelet coefficients, and the 32-channel scalograms are encapsulated into multi-time-step frames with sliding window.

In this work, after the CWT we further transform each channel signal's wavelet coefficients based time-scale representation into an energy based time-scale representation, namely '*scalogram*' [3], which can be obtained through Formula 3:

$$S(s) := (\int_{-\infty}^{+\infty} |Wf(s,u)|^2 du)^{\frac{1}{2}}. \quad (3)$$

As shown in Figure 3, it represents the distribution of the spectral energy in a signal, the hotter the pixel's color the more concentrate of energy in specific frequency range. The advantage of scalogram including two aspects: Firstly, the scalogram reflects detailed change of spectral information in both time and scale, and the spectral energy oscillations have been recognized as an indicator of various cognitive processes [24]. Secondly, each element of the scalogram is the percentage of spectral energy that the corresponding frequency component carries and the sum of all the elements is equal to 1. Therefore, the numerical range is naturally suitable for processing by '*Artificial Neural Networks (ANN)*'.

The sequential frames of one trial are constructed based on those scalograms, detailed in the following subsection.

2) Frame Construction: The frames are constructed after the scalograms of multi-channel signals have been obtained. Each frame is a grid-like $C \times S$ structure, which represents the spectral energy distribution in the C channels and the S selected scales within a time window. The procedure of constructing a frame can be summarized into four steps:

- Firstly, we should determine the length of the time window that a frame represents. For example, if we set the time window as 1s long with no overlap between

adjacent windows, then we can get 60 frames for a 60s long trial.

- Secondly, for each channel signal we add up the elements of the scalogram within the window along the time dimension. Then we can get a vector $v \in \mathbb{R}^S$ which represents the energy distribution in the scales within current window.
- Thirdly, we stack the obtained vectors of C channels together to construct the 2D frame in current window.
- Finally, we slide the window right with a window's length and repeat the step from 1 to 3 until all of the frames of one trial have been constructed.

3) Scale Selection: In order to reduce the computing burden, we also adopt '*Energy to Shannon Entropy Ratio (EER)*' to select some of the most representative scales. The optimal scales are selected when its spectral energy is high meanwhile its Shannon entropy is low. The criteria is presented as Formula 4:

$$r(s) = \frac{Energy(s)}{Entrophy_{sh.}(s)} \tag{4}$$

The energy of the scale '*s*' can be calculated through the sum of the energy that the '*n*' wavelet coefficient of this scale carries, as Formula 5:

$$Energy(s) = \sum_{i=1}^{n} |C_i(s)|^2 \tag{5}$$

The Shannon entropy describe the uncertainty of the energy distribution in scale '*s*'. The lower the entropy the more information the specific scale contains, as Formula 6:

$$Entropy_{sh.}(s) = -\sum_{i=1}^{n} P_i \log P_i \tag{6}$$

where the P_i is the probability distribution of the energy of the coefficient C_i in scale '*s*' and $\sum_{i=1}^{n} P_i = 1$, as Formula 7:

$$P_i = \frac{|C_i(s)|^2}{Energy(s)} \tag{7}$$

We calculated the EER for all channel signals according to the method mentioned above. After averaging the results we got the average EER for each scale, as shown in Figure 4. For example, we can select components in scale from 7 to 38 whose ratio is relatively high to construct the frames.

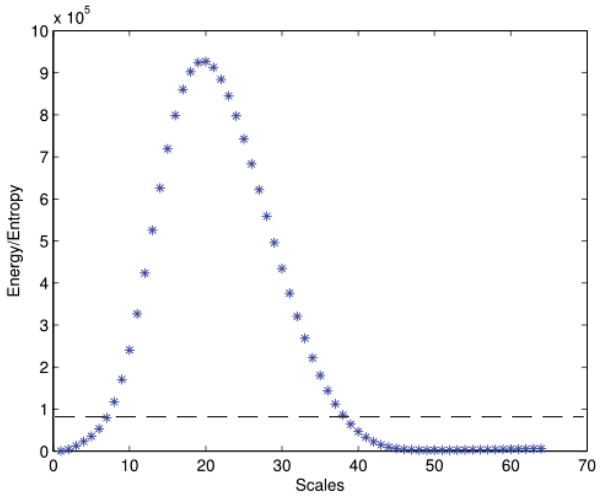

Fig. 4: The average Energy-Entropy Ratio is calculated for all channel signal, and the scale from 7 to 38 are selected for constructing the frames.

B. The Convolutional Recurrent Neural Network

Besides the preprocessing method mentioned above, we also propose a hybrid deep learning model, which is called the '*Convolutional Recurrent Neural Networks (C-RNN)*' by us, to conduct emotion recognition tasks. As shown in Figure 5, the model is a composition of two kinds of deep learning structures. It combines the powerful ability of the CNN in processing data with grid-like topology and the RNN in processing sequential data. The CNN unit works for mining cross-channel correlation and extracting features from the 2D frames. The '*Long Short-term Memory (LSTM)*' unit, which is a refined RNN structure, models the context information for long-term sequences that have arbitrary length. This hybrid model is quite suitable for processing two-dimensional sequential data.

The function of the C-RNN in this paper can be formulated as: $< M_1, ..., M_t, ..., M_n > \mapsto l$, the l represents the predicted emotion label of the frame sequence. The difference between our model and traditional '*many-to-one (M2O)*' model lies in how the predicted label generated. The label of traditional M2O model is determined by its last step's output, while the label of our model is determined by each step's output.

The average layer (decision layer) of the C-RNN is used for recording decision information in each time step, which is the basis for the final decision of the entire trial. Specifically, the final l is obtained by averaging the value in the softmax nodes of the decision layer in each time step. Then, the node with the maximum average probability determines which class of this trial belongs to, formulated as: $l = argmax(\frac{1}{n}\sum_{i=1}^{n} y_i)$. This strategy complies with our assumption that the participants' emotional rating is based on the entire experience in a trial.

The weights of the time distributed CNN are tied across time, so it can also be regarded as only a single CNN exists in this time-series model, and the convolutional filter size is deliberately designed for mining the correlation among different channels as well as scales.

The model is constructed and trained through some open source deep learning libraries, such as Keras[†] [6] . We next introduce the two components CNN and RNN of our hybrid model, respectively.

1) Convolutional Neural Networks (CNN): The Convolutional Neural Networks is a successful case of introducing findings in neuroscience to deep learning researches. It has achieved great success not only in the field of computer

[†]The implementation of preprocessing method and configuration of the model framework can be found on the web site: https://github.com/muzixiang/Multichannel_Biosignal_Emotion_Recognition

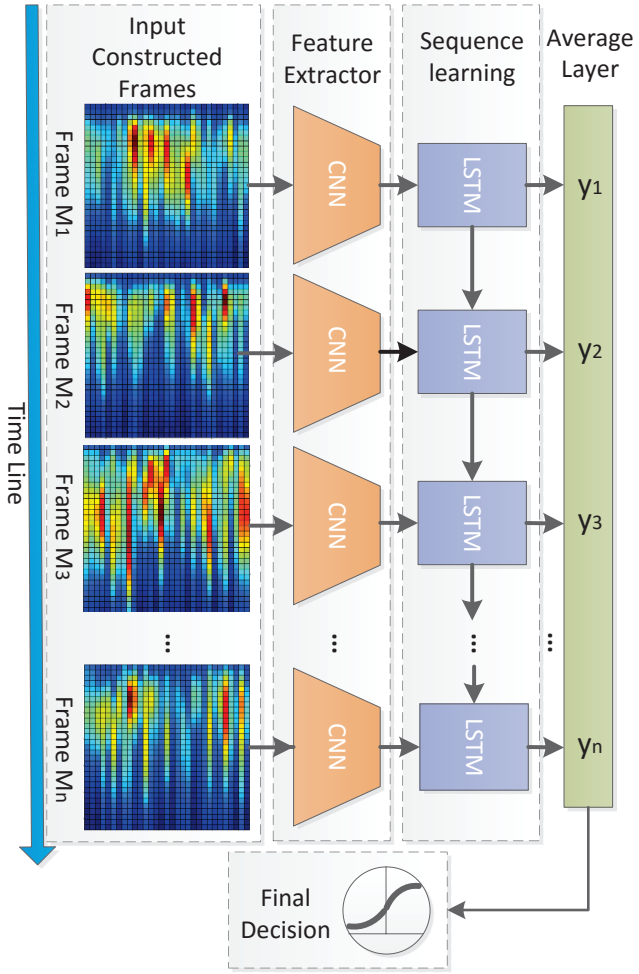

Fig. 5: The unfolded time chain form of the C-RNN model in emotion recognition. Each time step a frame is fed into the model, the CNN is responsible for extracting cross-channel correlation features through deliberately designed convolutional filters and the extracted information is further fed into the LSTM unit for context learning. Finally the decision layer give the recognition results based on the sequences of the entire trial.

vision but also in the fields of speech recognition and natural language processing, etc. The architecture and mechanism of CNN provides the possibility for neural networks in processing data with grid-like structure. The designed convolutional filters help for extracting multiple kinds of features automatically.

Generally speaking, a CNN is composed of one or several stacked convolutional layers. Each convolutional layer typically includes three processing stages, namely convolution stage, detector stage and pooling stage [14]. The convolutional stage is a process of applying convolutional filters to original 2D data. After this process, multiple feature maps are acquired from the input. The characteristics of the convolution stage includes sparse connectivity and parameter sharing. The mechanism of parameter sharing greatly decrease the amount of weight parameters in traditional full-connected neural networks, and in turn reduces the demand for parameter storage. The following detector stage is a non-linear transformation (e.g., a Sigmoid or ReLU activation function) of the obtained output from convolution stage. The last stage is another operation called pooling (e.g., Max Pooling and Average Pooling) which is a summary statistics of nearby results after detector stage, this stage helps the representation to be invariant to translation of input, and meanwhile the size of the input to next convolutional layer or a fully-connected layer can be reduced greatly.

In this work, we adopt two stacked convolutional layers as the basic structure of the CNN, the convolutional filter in the first convolutional layer was specifically set as $C \times 1$ with the purpose of mining cross-channel correlation information in a specific scale. After repeated convolutional operation along the scale dimension we obtain a feature map that contains S different cross-channel correlation information mined in the S scales, and we set the number of convolutional filters as 8 in this stage to extract 8 different kinds of correlation information, namely 8 different feature maps. Following the first convolutional layer is an average pooling layer with pooling size of 2×1 for aggregating correlation information in adjacent scales. After the first pooling stage, the size of a feature map is down sampled from $S \times 1$ into 16×1. The second convolutional layer is set as 16 different filters with size of 1×1, this setting helps to further fuse a specific scale range's information from prior 8 feature maps. Therefore, we get 16 feature maps, and each represents the fused information in different scale range. Similar to the first convolutional layer, we add an average pooling stage after this convolutional layer for information aggregation. Before connecting to the LSTM unit, a flatten operation needed for transforming the final 16 feature maps into an one-dimensional vector.

2) Recurrent Nueral Networks (RNN): After combining with the RNN unit, the hybrid model acquires the ability of learning a time series. The RNN is good at sequential modeling that traditional deep neural network (DNN) can't do well. The difference between the RNN and the DNN is the weights parameters are reused at every time step, so the number of parameters will not increase in proportion with the length of the input sequence. The RNN's ability relies on its recurrence structure, which can model context information from sequences with equal or different length. It is very important when we do not know which moment plays the most important role in the subject's final evaluation of the specific emotion they experienced in a trial.

The simple RNN's practical application has been hampered by its special design for difficult training, the RNN must faces the mathematical challenge of '*gradient vanish or explode*' in back propagation when its dependencies is too long [2]. Therefore, in order to reduce the difficulties in learning a long-term dependencies, some rectified recurrent units have been adopted to replace the usual units of the traditional RNN, including GRU and LSTM that combine '*gate*' mechanism in their structure. The gate can forget the information has been used and the self-loop structure allows the gradient to flow for long durations. These gated RNNs have gained great success in tasks of handwriting recognition, speech recognition, machine translation, image caption, parsing, etc [9].

In this work, we adopt LSTM as the RNN unit. As shown

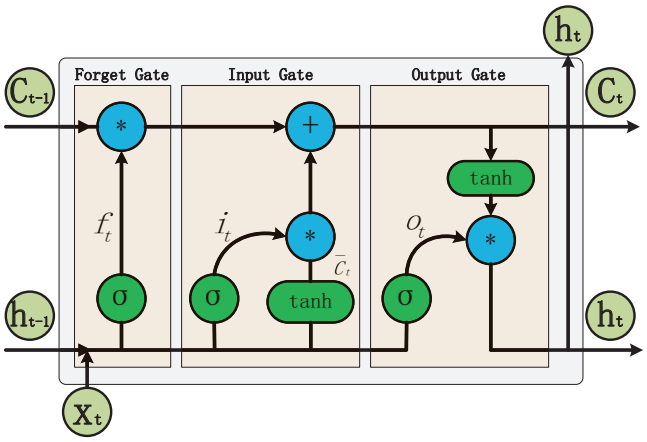

Fig. 6: The detailed structure of the LSTM unit, its function for context and sequence learning is based on those three gate mechanisms.

in Figure 5, the recurrence of a LSTM based RNN can be presented and comprehended through unfolding it into a chain form, each chain represents a time step in processing the data that output from the CNN. The cell states flows along with those chains, each chain has three gate structures that determine what information from prior step should be forgot and what information in current time step should be added into the main flow. A typical LSTM unit's structure is illustrated in Figure 6 and the mechanism of the gates is described as follows:

The first one is the '*Forget Gate*', which determines what information from the past should be forgot. The hidden state h_{t-1} from the prior LSTM cell and the current step's input x_t are concatenated into a new vector, after multiplication with the weight parameters W_f of the gate, each element's value of the output vector f_t is scaled between 0 and 1 through element-wise sigmoidal operation σ. This output f_t acts as decision vector, it helps to determine what information in the prior cell state C_{t-1} should be reserved through element-wise multiplication $C_{t-1} * f_t$. The '0' element causes the corresponding information in C_{t-1} will be wiped out, while the '1' means the corresponding information is allowed passing through. The output f_t of the gate is formalized as Equation 8.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{8}$$

The second one is the '*Input Gate*', the fulfillment of its function needs cooperation of two parallel layers. The tangent layer outputs candidate information $\bar{C}_t$ for selection, while the sigmoidal layer acts just as the forget gate, it decides what candidate information will be selected by outputting a decision vector i_t. After the element-wise multiplication of the candidate information by the decision vector $\bar{C}_t * i_t$, the final updating information that should be added to the cell state is determined. The tow layers' function is formalized as Equation 9 and 10, respectively.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{9}$$

$$\bar{C}_t = tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{10}$$

Therefore, the cell state C_t of the current chain is a combination of the reserved historical information of C_{t-1} and the updating information selected from $\bar{C}_t$, as Equation 11.

$$C_t = C_{t-1} * f_t + \bar{C}_t * i_t \tag{11}$$

The last one is the '*Output Gate*'. In a word, it decides outputting what hidden state h_t in current chain through multiplication of the decision vector o_t by the candidate information selected from C_t, as shown in Equation 12 and 13, respectively.

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{12}$$

$$h_t = tanh(C_t) * o_t \tag{13}$$

In a word, the LSTM based RNN in this model is used for learning contextual information from the feature sequence that extracted through the front CNN, and the recognition of emotional states for the entire trial is decided based on the output of LSTM in each time step.

III. Experiment and Discussion

In this section, we will show the effectiveness of our methods, and we also compare our method with two self-designed baselines and some peer-reviewed baselines.

A. Experimental Dataset

Our model was validated on the publicly available DEAP dataset [13], which includes multi-channel neurophysiological signals collected from 32 subjects. The subjects' various emotions were stimulated though 40 music videos that corresponding to different emotional genre. One stimuli is presented in one trial, and the signals were continually recorded during those trials. After each trial, the subjects subjectively evaluate their emotional experience on a two-dimensional emotional space which is proposed by Russell [19], where the two dimensions are Arousal (it ranges from relaxed to aroused) and Valence (it ranges from pleasant to unpleasant) respectively. The ratings are continuously distributed from 1 to 9 in each dimension, we divide and label the trials into two classes for Valence and Arousal respectively (pleasant: > 5, unpleasant: ≤ 5; aroused: > 5, relaxed: ≤ 5).

For the sake of the limitation of the number of trials in the dataset for training our model, we need to adopt some data augmentation strategies before training to avoid overfitting. In this paper, we choose to tackling this problem through adding gaussian white noise with 5dB SNR (Signal to Noise Ratio) to the original channel signal, finally each channel signal is augmented 250 times to make us have 10,000 sequences per subject. Such enough training data not only helps the model to tolerate the inherent noise underlies the physiological signal but also helps the model with large number of parameters converge and generalize well.

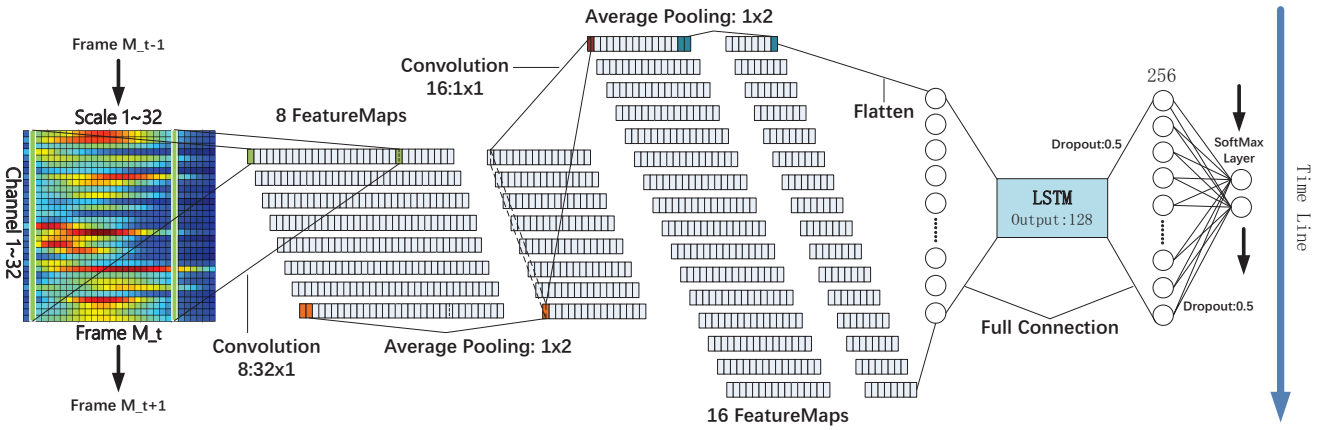

Fig. 7: The settings and mechanism of C-RNN in processing multi-channel EEG frames and conducting emotion recognition.

B. Experimental Settings

The structure and the detailed settings of hyper parameters for the C-RNN model is illustrated in Figure 7. Notice that we adopt '*dropout*' in the last two layers after LSTM to prevent overfitting [20].

We use the stratified K-folds cross validation methods to evaluate the performance of our approach. It can give us a relatively impartial evaluation of our model, where each fold contains approximately the same percentage of trials of each class as the whole set. We set the K as 5 and average the performance of the 5-folds validation processes as the final results reported in Figure 8.

For comparison, we select the Random Forest (RF) and Support Vector Machine (SVM) classifiers implemented in Scikit-learn toolkit with multiple hand-engineered features as the baseline approaches. The hand-engineered features includes 9 linear features as well as 3 nonlinear features extracted from the signals. These features are the signal's peak-to-peak mean value, the variance value, the root mean square value, three hjorth parameters (complexity, mobility and activity) of the amplitude, the max power spectral and its corresponding frequency, the sum of power spectral, the c0 complexity, the correlation dimension and the shannon entropy. Therefore, the dimension of the combined feature vector for one trial is 384 (12 hand-engineered features $\times$ 32 channels). To make sure the performance between our model and the baseline methods can be compared, we apply the same 5-folds division on SVM and RF based methods.

C. Results and Discussion

The average accuracy of our method and baselines are reported in Figure 8. Since we focus on the trial-oriented recognition rather than some segment-oriented recognition tasks, therefore here we only select those relevant peer-reviewed studies out for comparison. Chen el al., (2015) extracted over a thousand of features and studied different feature selection method and adopted Hidden Markov Models (HMM) to perform trial-oriented emotion recognition on a subset of the total 32 participants [5]. Rozgic et al., (2013) proposed a segment-level feature extraction and four different fusion strategies for constructing trial-level features and conducted final recognition of emotional states through K-PCA and RBF-SVM for the trial [17]. Koelstra et al., (2012) simply extracted trial-level features as our self-designed baselines and performed recognition through naive Bayes classifier [13].

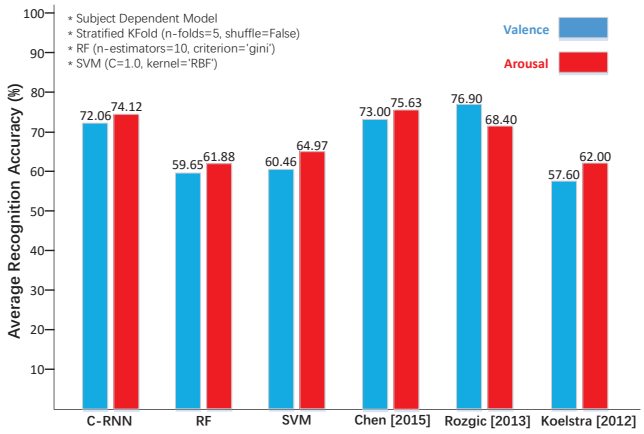

Fig. 8: Performance comparison between relevant methods, the sliding window width of our method is set as 1s long.

The comparison shows the effectiveness of our method. The performance of our method is close to that of the method adopted by Chen et al. (2015). Considering the characteristics of the HMM that each step's output is only related to a few prior states and the parameters are not shared between steps, the RNN based method is a good substitute for it when the physiological signal sequence is quite long and has variable length. The performance on Valence dimension in work of Rozgic et al., (2013) is relatively high, but this method shows a degraded performance on Arousal dimension, the proposed segment-level to trial-level feature projection methods is time consuming when dealing with large quantities of samples. The scenarios of using this method and the HMM based method is when the whole data set has been provided with. They are not suitable in performing incremental learning (the samples is continuously gathered and fed into the model) that the deep

learning is good at.

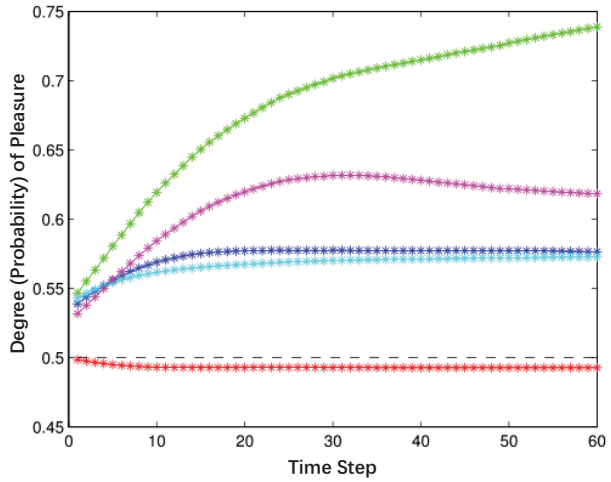

Fig. 9: The realtime prediction on Valence for five selected 60s trials. The results are obtained through storing each step's output probability. The probability greater than 0.5 indicates a pleasant experience, while the probability lower than 0.5 indicates an relatively unpleasant experience.

As shown in Figure 9, besides the ability in classifying the entire trial, this RNN based model also has the ability in realtime prediction for each time step. However, the trials in the DEAP dataset do not have real time labels for each time step. We can not validate the accuracy of the predictions currently.

IV. Conclusions

In this paper, we have proposed a hybid deep learning model, C-RNN, which integrates CNN and RNN , for emotion recognition based on multi-channel EEG signals. Specifically, the CNN component has the ability in mining inter-channel or cross-modal correlation information. On the other hand, the RNN (i.e., LSTM) based model structure can learn long-term dependencies and contextual information from sequences. Practically, instead of manually designing task related features as in traditional approaches, we propose a novel pre-processing method that transforms the multi-channel EEG data into a 2D frame representation. The proposed method has been shown effective in the trial-level emotion recognition task. It also has a potential in giving predictions not only for an entire trial but also for each time step, which is very important in realtime emotion monitoring scenarios. One problem that we need to address in the future is the influence of insufficient training data on the performance of our hybrid DL model.

Acknowledgment

This work is funded in part by the Chinese National Program on Key Basic Research Project (973 Program, grant no.2013CB329304 and 2014CB744604), the Chinese 863 Program (grant no.2015AA015403), the Natural Science Foundation of China (grant no.61272265 and 61402324), and the Tianjin Research Program of Application Foundation and Advanced Technology (grant no.15JCQNJC41700).

References

[1] American Psychiatric Association. *Diagnostic criteria from DSM-IV-tr*. American Psychiatric Pub, 2000.

[2] Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166, 1994.

[3] V. J. Bolós and R. Benítez. The wavelet scalogram in the study of time series. In *Advances in Differential Equations and Applications*, pages 147–154. Springer, 2014.

[4] W. B. Cannon. The james-lange theory of emotions: A critical examination and an alternative theory. *The American journal of psychology*, 39(1/4):106–124, 1927.

[5] J. Chen, B. Hu, L. X. Xu, and et al. Feature-level fusion of multimodal physiological signals for emotion recognition. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 395–399. IEEE, 2015.

[6] F. Chollet. Keras: Deep learning library for theano and tensorflow. https://github.com/fchollet/keras, 2015.

[7] R. N. Duan, J. Y. Zhu, and B. L. Lu. Differential entropy feature for eeg-based emotion classification. In *6th International IEEE/EMBS Conference on Neural Engineering (NER)*, pages 81–84. IEEE, 2013.

[8] T. Gandhi, B. K. Panigrahi, and S. Anand. A comparative study of wavelet families for eeg signal classification. *Neurocomputing*, 74(17):3051–3057, 2011.

[9] A. Graves. Supervised sequence labelling with recurrent neural networks. In *Neural Networks*, pages 15–35. Springer, 2012.

[10] D. Huang, C. T. Guan, K. K. Ang, and et al. Asymmetric spatial pattern for eeg-based emotion detection. In *The 2012 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2012.

[11] M. S. Hussain, R. A. Calvo, and P. A. Pour. Hybrid fusion approach for detecting affects from multichannel physiology. In *International Conference on Affective Computing and Intelligent Interaction*, pages 568–577. Springer, 2011.

[12] S. Jirayucharoensak, S. Pan-Ngum, and P. Israsena. Eeg-based emotion recognition using deep learning network with principal component based covariate shift adaptation. *The Scientific World Journal*, 2014(627892), 2014.

[13] S. Koelstra, C. Muhl, and M. Soleymani. Deap: A database for emotion analysis using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2012.

[14] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[15] S. Marsella, J. Gratch, and P. Petta. Computational models of emotion. *A Blueprint for Affective Computing-A sourcebook and manual*, 11(1):21–46, 2010.

[16] P. C. Petrantonakis and L. J. Hadjileontiadis. Emotion recognition from eeg using higher order crossings. *IEEE Transactions on Information Technology in Biomedicine A Publication of the IEEE Engineering in Medicine & Biology Society*, 14(2):186–97, 2010.

[17] V. Rozgić, S. N. Vitaladevuni, and R. Prasad. Robust eeg emotion classification using segment level decision fusion. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1286–1290. IEEE, 2013.

[18] R. Rubinstein, A. M. Bruckstein, and M. Elad. Dictionaries for sparse representation modeling. *Proceedings of the IEEE*, 98(6):1045–1057, 2010.

[19] J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.

[20] N. Srivastava, G. E. Hinton, A. Krizhevsky, and et al. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[21] C. J. Stam. Nonlinear dynamical analysis of eeg and meg: review of an emerging field. *Clinical Neurophysiology*, 116(10):2266–2301, 2005.

[22] A. Subasi. Automatic recognition of alertness level from eeg by using neural network and wavelet coefficients. *Expert systems with applications*, 28(4):701–711, 2005.

[23] M. Valstar, B. Schuller, K. Smith, and et al. Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 3–10. ACM, 2013.

[24] L. M. Ward. Synchronous neural oscillations and cognitive processes. *Trends in cognitive sciences*, 7(12):553–559, 2003.