Efficient Modeling of Long Temporal Contexts for Continuous Emotion Recognition

Jian Huang
National Laboratory of Pattern Recognition, (NLPR)
Institute of Automation, CAS,
School of Artificial Intelligence
University of Chinese Academy of Sciences
Beijing, China
jian.huang@nlpr.ia.ac.cn

Jianhua Tao
National Laboratory of Pattern Recognition, (NLPR)
Institute of Automation, CAS,
School of Artificial Intelligence
University of Chinese Academy of Sciences,
CAS Center for Excellence in Brain Science and Intelligence Technology,
Institute of Automation, CAS
Beijing, China
jhtao@nlpr.ia.ac.cn

Bin Liu
National Laboratory of Pattern Recognition, (NLPR)
Institute of Automation, CAS, School of Artificial Intelligence
University of Chinese Academy of Sciences
Beijing, China
liubin@nlpr.ia.ac.cn

Zhen Lian
National Laboratory of Pattern Recognition, (NLPR)
Institute of Automation, CAS, School of Artificial Intelligence
University of Chinese Academy of Sciences
Beijing, China
lianzheng2016@ia.ac.cn

Mingyue Niu
National Laboratory of Pattern Recognition, (NLPR)
Institute of Automation, CAS, School of Artificial Intelligence
University of Chinese Academy of Sciences
Beijing, China
niumingyue2017@ ia.ac.cn

***Abstract*—Continuous emotion recognition is a challenging task due to its difficulty in modeling long-term contexts dependencies. Prior researches have exploited emotional temporal contexts from two perspectives, which are based on feature representations and emotional models. In this paper, we explore the model based approaches for continuous emotion recognition. Specifically, three temporal models including LSTM, TDNN and multi-head attention models are utilized to learn long-term contexts dependencies based on short-term feature representations. The temporal information learned by the temporal models allows the network to more easily exploit the slow changing dynamics between emotional states. Our experimental results demonstrate that the temporal models can model emotional long-term dynamic information effectively. Multi-head attention model achieves best performance among three models and multi-model combination models further improve the performance of continuous emotion recognition significantly.**

Keywords—continuous emotion recognition, temporal model, TDNN, multi-head attention, multi-model combination model.

I. INTRODUCTION

To enable the human-machine interfaces more harmoniously, it is bound to have emotional intelligence [1]. Automatic emotion recognition is key factor for building intelligent human-machine interfaces that can adapt to the affective state of the user. Given the initial stereotypical expressions, most of the works focused on modeling an emotional space consisting of discrete basic states such as anger, disgust, fear, happiness, sadness, and surprise [2]. However, human expresses emotional state related information through numerous subtle ways. A person's emotional state can be described by continuous space, which uses numerical values to indicate the type and degree of the emotions continuously [3], such as arousal (calm versus active) and valence (negative versus positive). Continuous emotion model has the advantages to represent multiple subtle and complex emotional states.

Emotional state is a gradual and smooth process with temporal change. A continuous space not only allows for a more complete description of a complex emotional state [4], but also leads itself better to continuous tracking and temporal modeling. Therefore, it is essential to model long-term context dependencies when making frame-level dense prediction for continuous emotion recognition. Modeling the temporal dynamics to capture the long-term dependencies between emotional behaviors, requires an emotional model which can effectively deal with long temporal contexts. Two types of approaches to exploit temporal contexts are using feature representations, which are designed to present this information to the model in a suitable form, or using emotional models, which can learn the long-term dependencies based on short-term feature representations.

Many researchers have investigated this problem using feature representations. Valstar et al. [5] showed that it was necessary to consider larger window to obtain segment-level features e.g. four seconds for arousal and six seconds for valence. Le et al. [6] and Cardinal et al. [7] found that increasing the number of contextual frames with deep neural network (DNN) could improve the performance. Povolny et al. [8] applied simple frame stacking and temporal content summarization to consider contextual information. Huang et

978-1-7281-3888-6/19/$31.00 ©2019 IEEE

al. [9] utilized temporal pooling subsampling to obtain wider context information to achieve emotional temporal modeling.

On the other hand, many emotional models have been explored to learn emotional dynamic information. Nicolaou et al. [10] exploited the temporal dependencies over a dimensional domain by extending the relevance vector machine regression framework, to capture the output structure and the covariance within a predefined time window. Recurrent neural networks (RNNs) and Long short-term memory (LSTM) have been shown to achieve state-of-art performance in sequence-to-sequence modeling. Wöllmer et al. [11] first proposed a method based on LSTM for continuous emotion recognition that included modeling of long-range dependencies between observations. Zhang et al. [12] utilized RNN to learn spatial and temporal dependencies for emotion recognition.

Many researchers have utilized Convolutional Neural Network (CNN) based methods to capture long-term temporal dependencies of emotion recognition. Khorram et al. [13] investigated two convolutional network architectures, dilated convolutional networks and downsampling/upsampling networks for capturing long-term temporal dependencies. The results achieved good performance and generated smooth output trajectories on the RECOLA dataset. Li et al. [14] proposed a CNN with two different groups of filters to capture both temporal and frequency domain context information for emotion recognition. Most works utilize 2D CNN for emotion modeling, which means that frequency domain features are employed for speech signals and single frame image is processed for video data. Recently, there is a recent trend in deep system design which attempts to derive features of the input signal directly from raw unprocessed input signals. Huang et al. [15] utilized 3D CNN to directly learn from video data for end-to-end continuous emotion recognition. Trigeorgis et al. [16] trained a convolutional recurrent network for continuous emotion recognition using the time domain signals directly. Due to single dimension of speech signals, they employed temporal CNN which is 1D convolutional neural network. Similar works are presented in EnvNet [17] and SoundNet [18].

Actually, the network structure of temporal CNN is similar to time-delay neural network (TDNN). TDNN [19] is another neural network model with the capability of capturing the dynamic relationship between consecutive observations. TDNNs have been shown to effectively learn the temporal dynamics of the signal for speech recognition [20]. Meng et al. [21] proposed a two-stage architecture that combines a simple regression algorithm and TDNN for automatic continuous affective state prediction from facial expressions. The experimental results demonstrated that the use of a two-stage approach combined with the TDNN, to take into account previously classified frames, significantly improves the overall performance of continuous emotional state estimation. Sarma et al. [22] tried interleaves TDNN-LSTM with time-restricted self-attention and achieving a weighted accuracy of 70.6% in IEMOCAP. These results reveal the availability of TDNN to model emotional temporal contexts.

The above models are either based on RNNs or CNNs. Recently, Vaswani et al. [23] proposed a no-recurrence sequence-to-sequence Transformer model to achieve state-of-the-art performance on machine translation, which also achieved most promising performance in speech recognition [24]. Its fundamental module is self-attention, a mechanism relates all the position-pairs of a sequence to extract a more expressive sequence representation. Since the self-attention can draw the dependencies between different positions through the position-pair computation rather than the position-chain computation of RNNs, it just needs to be calculated once to obtain the transformed representation. Therefore, the Transformer model also can learn long temporal dependencies on the longer span of time.

The methods discussed above make use of different emotional models that are able to capture temporal information. Inspired by their works, we utilize LSTM, TDNN and multi-head attention models to model long temporal contexts for continuous emotion recognition. In the following, Section 2 briefly introduces the proposed models. Section 3 presents the database and feature sets. Section 4 describes experimental results and analysis. Section 5 concludes this paper.

II. Proposed Methods

In this paper, we utilize three temporal models to learn long-term contexts dependencies of continuous emotion. In this section, we firstly describe the TDNN and multi-head attention models briefly except the LSTM model due to its universality. Then, we introduce individual emotional temporal models and their combination models.

A. TDNN

TDNN is a fully-connected forward-feedback neural network model with delays in the nodes of the hidden and output layers. In a TDNN, different layers or sets of layers can act on different time scales. As such, it can be seen as a type of CNN operating over the time dimension. Particularly, current input signal is augmented with delayed copies of the previous input values. The neural network is time-shift invariant since it has no internal state, as shown in Fig. 1. Besides, the first few layers look at smaller time scales and produce more abstract higher level features. The later layers take larger time windows over these abstract features as the inputs. In term of emotion recognition, this means that an instant of an emotional expression is recognized by taking into account not only the input features describing that instant, but also the input features describing the previous instants, i.e., how the expression evolved over time to the current state. We utilize TDNN to capture the temporal relationship between the predictions on continuous emotional states.

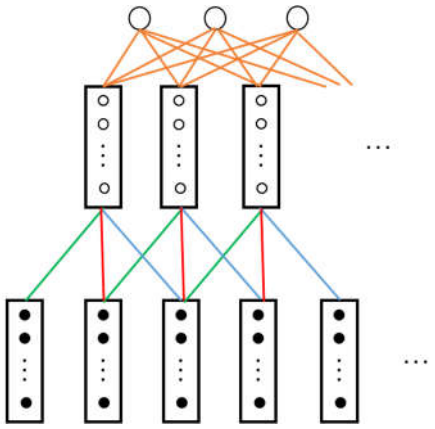


Fig. 1. Overall architecture of the TDNN.

B. Multi-head Attention

The multi-head attention module is a core module of the Transformer network, which plays an essential impact of learning long temporal dependencies. It is based on self-attention module to extract a more expressive sequence representation. The multi-head attention module leverages different attending representations jointly for emotion modeling. Besides, it extends the conventional attention mechanism to have h multiple heads, where each head can generate a different attention distribution. This allows each head to have a different role on attending the representations.

Specifically, the multi-head attention calculates h times Scaled Dot-Product Attention in Figure 2. Before performing each attention, there are three linear projections to transform the queries Q, keys K and values V to more discriminated representations respectively. Then, each Scaled Dot-Product Attention is calculated independently, and their outputs are concatenated and fed into another linear projection to obtain the final outputs:

$$\text{MultiHead}(Q,K,V) = \text{Concat}(\text{head}_1,\dots,\text{head}_h)W^O$$

$$\text{head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right)$$

where the projection are parameter matrices.

Scaled Dot-Product Attention has three inputs: queries, keys of dimension d_k and values of dimension d_v. One query's output is computed as a weighted sum of the values, where each weight of the value is computed by a designed function of the query with the corresponding key. The dot products of the query with all keys, divided each by $\sqrt{d_k}$ and applied a softmax function to obtain the weights on the values.

$$\text{Attention}(Q,K,V) = softmax\left(QK^T/\sqrt{d_k}\right)V$$

where the scalar $1/\sqrt{d_k}$ is used to prevent softmax function into regions that has very small gradients.

C. Emotional Temporal Modeling

We build continuous emotion recognition systems to model long-term contexts dependencies with temporal models described in previous section. Firstly, we utilize single temporal models for emotion modeling as shown in Fig. 4. The inputs are emotional features extracted from different modalities. The outputs are the predictions corresponding to every frame. The emotional temporal model is responsible to modeling emotional long temporal contexts dynamic information. In addition, the purpose of initial linear layer is to transform original emotional features to common emotional feature space, and final linear layer is to convert high level representations to continuous emotion space and get the predictions.

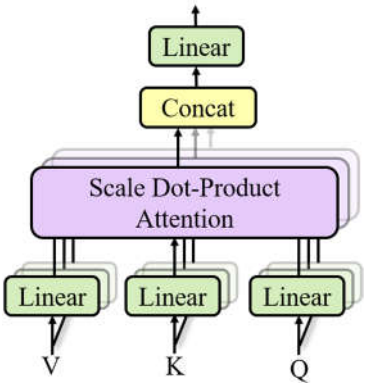


Fig. 2. Multi-head attention consists of several attention layers running in parallel.

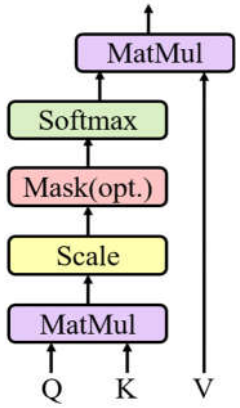


Fig. 3. Scaled Dot-Product Attention.

Three models have their own advantages in emotional temporal modeling. TDNN can process window temporal contexts with delayed copies of the previous input values. LSTM can learn long-term dynamic information since their outputs are influenced by the outputs of hidden layer involving previous information and the current input. Multi-head attention model can achieve long span modeling from global information with self-attention mechanism. These three models employ temporal convolution, recurrence and attention mechanism respectively. On basis of them, we build sequence-to-sequence continuous emotion recognition systems respectively.

Next, various temporal models are combined, called multi-model combination models, to strengthen the ability of learning emotional temporal contexts information. We establish emotion recognition systems based on pairwise combination, as shown in Fig. 5(a)(b)(c). Fig. 5(a) shows the combination of TDNN and LSTM represented by "TDNN+LSTM". TDNN learns short term dependencies from the emotional features to produce high level features, and LSTM is responsible to learn long temporal contexts from high level features. We take the encoder part of original Transformer model to perform continuous emotion recognition. The common structure of encoder part is composed of multi-head attention module followed by one feed forward layer which is fully connected layer [23]. We replace the fully connected layer with TDNN layer to combine the multi-head attention and TDNN models, as shown in Fig. 5(b) represented by "Attention+TDNN".

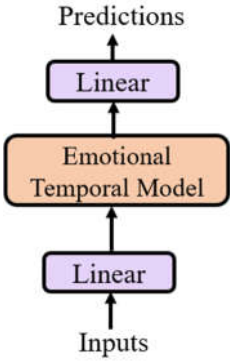
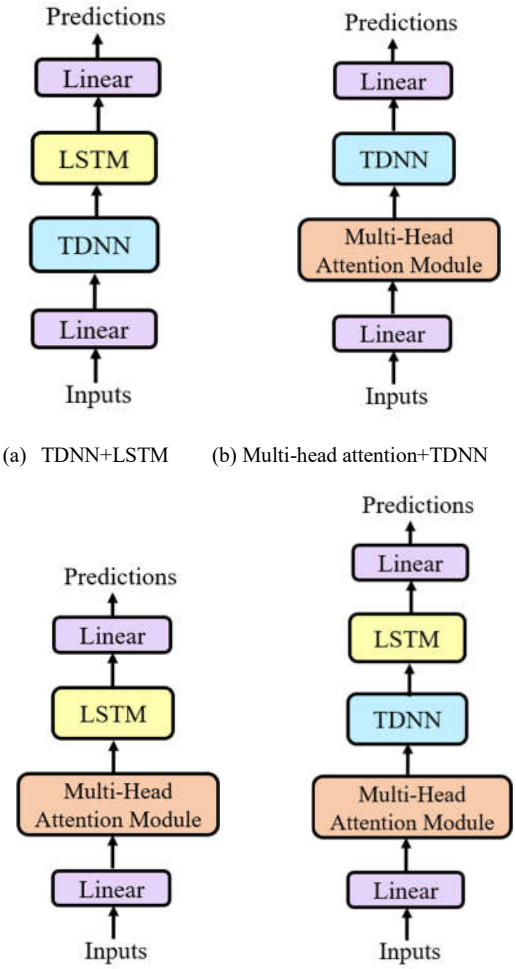


Fig. 4. Continuous emotion recognition system with single temporal model.



(a) TDNN+LSTM (b) Multi-head attention+TDNN

(c) Multi-head attention+LSTM +LSTM (d) Multi-head attention+TDNN +LSTM

Fig. 5. Continuous emotion recognition systems combining different emotional temporal models.

In Fig. 5(c), the feed forward layer is still fully connected layer and the multi-head attention module is followed by the LSTM layer, which is the combination of the multi-head attention and LSTM models represented by "Attention+LSTM". Finally, we add the LSTM layer to the back of the TNDD layer of Fig. 5(b) to combine three temporal models represented by "Attention+TDNN+LSTM", as shown in Fig. 5(d). The goal of multi-model combination is to absorb the advantages of different temporal models to better learn emotional long temporal contexts dependencies.

III. Database and Feature Sets

A. Database

In this study, we use Audio/Visual Emotion Challenge and Workshop (AVEC 2017) database based on Sentiment Analysis in the Wild (SEWA) [25] to show the benefits of our proposed methods. SEWA collects spontaneous and naturalistic human-human interactions in the wild consisting of audio, video and text modalities. The recordings are annotated time-continuously in terms of the emotional dimensions including arousal for the emotion activation, valence for the emotion positiveness and liking for the user's preference. The recording was based on the form of dyadic interactions and the conversations were asked to discuss the commercial product they had just viewed. The duration of each conversation is at most 3 minutes. All three emotional dimensions are annotated every 100ms and scaled into [-1, +1]. There are 64 German subjects in the dataset and are divided into training set with 36 subjects, development set with 14 subjects and testing set with 16 subjects. We focus on the estimation of arousal and valence in this work.

B. Feature Sets

In this work, we extract the emotional features from audio and visual modalities. The extended version of the Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) feature set [26] are adopted as the audio features. eGeMAPS is an expert-knowledge based feature sets consisting of 23 acoustic low-level descriptors (LLDs) extracted every 10 ms over a short-term frame, which have been applied in emotion recognition tasks successfully [5]. The LLDs set consists of energy, spectral and cepstral features, pitch, voice quality, and micro-prosodic features. The functionals including arithmetic mean and the coefficient of variation are computed on all LLDs. Segment-level acoustic features are computed over segments of 4 seconds. Overall, the acoustic baseline feature sets contain 88 dimensional features. The extraction of the LLDs and the computation of the functionals are done using the openSMILE toolkit [27].

The geometric features are adopted as visual features. We extract geometric features related to the position and expression of the subjects' face including face orientation, eye points and facial landmarks by Openface [28]. The 49 facial landmarks are aligned with a mean shape from stable points (located on the eye corners and on the nose region). Then, we compute the difference between the coordinates of the aligned landmarks and those from the mean shape, and also between the aligned landmark locations in the previous and the current frame. The same operations are applied for face orientation and eye points. We also computed the Euclidean distance between the median of the stable landmarks and each aligned landmark in a video frame. The geometric feature sets are interpolated by a piecewise cubic Hermite polynomial to cope with dropped frames. Finally, we apply a PCA to retain 55 dimensional features.

IV. Experiment and Analysis

A. Experimental Setup

All our experiments are done using TensorFlow. The implementation of TDNN is based on 1D convolutional neural network with conv1d layer. The output channels are 64 with 3 convolutional kernel. For single TDNN emotional model, we utilize a stack of 4 identical TDNN layers. The multi-head attention module is composed of a stack of 2 identical layers which are 4-head attention sublayer. And, each sublayer has a residual connection and layer normalization. The number of hidden nodes of multi-head attention sublayer are 64. For single LSTM emotional model, we use a stack of 2 LSTM layers with 64 hidden nodes.

For multi-model combination models, we combine different network structures to perform continuous emotion recognition. The parameters settings are similar to previous single models. The "TDNN+LSTM" model includes two TDNN layer and two LSTM layers. The "Attention+TDNN" model is composed of the multi-head attention module followed by one TDNN layer. The "Attention+LSTM" model includes the multi-head attention module followed by fully connected layer and one LSTM layer. We add one LSTM layer to the back of the "Attention+TDNN" model to achieve "Attention+TDNN+LSTM" model.

We use adam optimization algorithm [29]. The batch size is 3. The maximum training epochs are 70. The evaluation measure is the Concordance Correlation Coefficient (CCC) [30]. Due to no availability of the testing set, we utilize the training set to train models and the development set to evaluate the performance.

B. *Continuous Emotion Recognition with Single Temporal Models*

As discussed in the previous section, TDNN, LSTM and multi-head attention models are good candidates for real-time emotional state prediction at unit level. They capture the dynamic relationship existing between consecutive units of expressions. We utilize these three temporal models using audio and visual features to perform continuous emotion recognition in arousal and valence individually.

The experimental results are shown in Table I. Three temporal models can model emotional long-term dynamic information effectively and accomplish valid continuous emotion recognition results. We can observe that the performance of LSTM is better than TDNN both in arousal and valence. LSTM uses a dynamically changing contextual window over all of the sequence history due to recurrent structure resulting in better ability of long temporal contexts. TDNN only can cover local context with delayed inputs so that its design is better at short temporal modeling. Thus, LSTM can obtain better performance than TDNN. But, TDNN can achieve faster model training due to parallel computing compared with LSTM.

It is worth noticing that the performance of multi-head attention model is mostly better than LSTM except the audio modality in valence. Although the improved performance is not high, the multi-head attention model makes positive effects on continuous emotion recognition. The results reveal that multi-head attention mechanism, without recurrence and convolution structure, can model emotional long-term dynamic information effectively. The multi-head attention model completes the computation of self-attention mechanism based on global information on the long span of time. Thus, the multi-head attention model can learn longer span temporal contexts dependencies and improve the performance significantly. This indicates that the multi-head attention model has great potential to build more promising and robust continuous emotion recognition systems.

TABLE I. THE CCC PERFORMANCE OF DIFFERENT SINGLE TEMPORAL MODELS IN AROUSAL AND VALENCE FROM AUDIO AND VIDEO FEATURES.

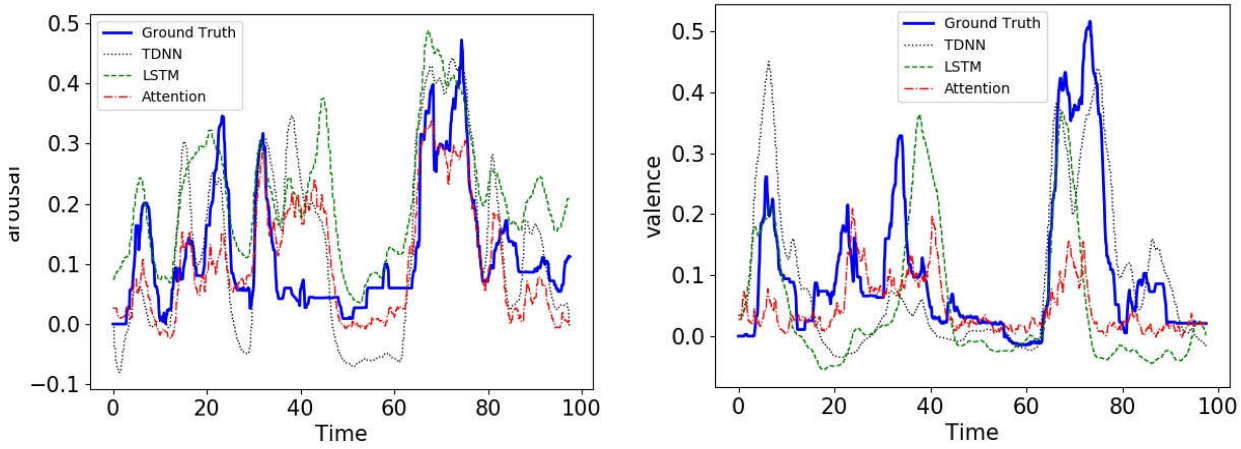
	Arousal		**Valence**	
	Audio	***Visual***	***Audio***	***Visual***
TDNN	0.385	0.551	0.421	0.508
LSTM	0.426	0.563	**0.451**	0.532
Multi-head attention	**0.459**	**0.581**	0.438	**0.554**

This also verifies the strong strength and universality of the multi-head attention model on sequence modeling. Therefore, incorporating long-term temporal dependencies is critical for continuous emotion recognition tasks. In addition, visual features achieve better performance than audio features in arousal and valence, which is similar to the works [9][25]. In audio features, the performance of arousal and valence is comparable while the performance of arousal is better than valence in visual features.

We take one sample of the development set to compare the effectiveness of three temporal models in arousal and valence, as shown in Fig. 6. The ground truth (the blue line) shows emotion evolves intensively in a short period of time, making it difficult to predict emotional dynamic precisely. The black, green and red one are the predictions of the TDNN, LSTM and multi-head attention models respectively. We can observe that the predictions of TDNN deviates the ground truth severely sometimes limited by the accessibility of local information. The predictions of LSTM and multi-head attention model can relieve this problem actually. However, these exists some lags and bias problems in the predictions of LSTM. The predictions of multi-head attention model are closest to the ground truth and depict the trend of emotional change. This reveals the superiority of multi-head attention model of emotion modeling. Efficient temporal modeling can promote the performance of continuous emotion recognition significantly.

C. *Continuous Emotion Recognition with Multi-model Combination Models.*

In order to better recognize continuous emotions, the crucial emotional temporal dependencies should be well modeled. In this work, multiple temporal models are combined to promote the ability of modeling long temporal



(a) Arousal (b) Valence

Fig. 6. The visualization of the predictions produced by the TDNN, LSTM and multi-head attention models against the ground truth.

contexts dependencies. We conduct the experiments based on multi-model combination models described in section II. The corresponding experimental results are shown in Table II.

We can observe that multi-model combination models can improve the performance in most of the situations. The "TDNN+LSTM" model achieve better performance than the single TDNN and LSTM models, benefited from the combination of short temporal and long temporal modeling. The performance of the "Attention+TDNN" is better than the "TDNN_LSTM" due to excellent ability of long temporal modeling of the multi-head attention model. However, the "Attention+TDNN" model is just optimization of the Transformer network, which can't take full advantage of the TDNN model. Thus, the performance improvement is limited. The "Attention+LSTM" model further improves the performance of emotion recognition systems, showing the superiority of LSTM over TDNN. Besides, the "Attention+LSTM" model achieves best performance in valence with audio features 0.477. The combination of the multi-head attention and LSTM model can model long temporal dependencies to improve the performance.

Next, we combine three temporal models together to learn long temporal contexts dependencies for continuous emotion recognition. The performance of the "Attention+TDNN+LSTM" model achieves best performance in arousal with audio features 0.519 and visual features 0.623. In addition, the "Attention+TDNN+LSTM" model achieves best performance in valence with visual feature 0.647. As for audio features, the multi-model combination models don't make difference in valence. The results show that multi-model combination models can achieve significant improvements in comparison with single temporal models. It's essential for continuous emotion recognition to consider long temporal contexts dependencies. Different from the results of single temporal models, the performance of arousal is better than valence in audio features while the performance of valence is better than arousal in visual features due to its higher promotion. The multi-model combination models can model longer temporal contexts of valence to improve the performance.

Finally, we compare our results with two other methods of the literatures. Ringeval et al. [25] utilized SVM as continuous emotion recognition model, which are the baseline work of AVEC 2017. Dang et al. [31] utilized Gaussian Mixture Regression model with audio features and Relevant Vector Machines model with visual features. The corresponding performance are listed in Table III. The features of the three models are same. The results show that our methods can improve the performance significantly compared with the baseline results. Furthermore, our methods achieve better performance than Dang's works. Therefore, our proposed models can efficiently model emotional long temporal contexts for continuous emotion recognition to improve the performance. The results verify the effectiveness of our proposed methods.

TABLE II. THE CCC PERFORMANCE OF DIFFERENT SINGLE TEMPORAL MODELS IN AROUSAL AND VALENCE FROM AUDIO AND VIDEO FEATURES.

	Arousal		**Valence**	
	Audio	***Visual***	***Audio***	***Visual***
TDNN+LSTM	0.461	0.591	0.431	0.585
Attention+TDNN	0.471	0.612	0.315	0.557
Attention+LSTM	0.506	0.616	**0.477**	0.634
Attention+TDNN+LSTM	**0.519**	**0.623**	0.421	**0.647**

TABLE III. CCC COMPARISON BETWEEN OUR METHODS AND OTHER METHODS

	Arousal		**Valence**	
	Audio	***Visual***	***Audio***	***Visual***
Ringeval et al. [25]	0.344	0.466	0.351	0.400
Dang et al. [31]	0.454	0.518	0.446	0.583
Our methods	0.519	0.623	0.451	0.647

V. CONCLUSION

This work explores different temporal models to learn long temporal contexts dependencies for continuous emotion recognition. TDNN can achieve short temporal modeling with delayed inputs, but can't learn long temporal contexts and obtain limited performance. LSTM is common emotional sequence modeling model and can achieve better performance. The multi-head attention model can model long temporal contexts dependencies to achieve best performance, which is based on self-attention mechanism to attend global information on the long span of time. The results show the potential benefits of the multi-head attention model to obtain more promising performance on continuous emotion recognition. The multi-model combination models can improve the performance significantly due to enhanced ability of temporal modeling. The combination of three models can achieve best performance in arousal and valence. Our methods also achieve better performance than other methods, verifying the effectiveness of our proposed methods. In the future work, we will explore better temporal models for continuous emotion recognition.

ACKNOWLEDGMENT

This work is supported by the National Key Research & Development Plan of China (No.2017YFC0822502), the National Natural Science Foundation of China (NSFC) (No.61425017, No.61831022, No.61773379, No.61771472), and the Strategic Priority Research Program of Chinese Academy of Sciences (No.XDC02050100).

REFERENCES

[1] J. Tao, T. Tan, "Affective computing: A review," International Conference on Affective Computing and Intelligent Interaction, pp. 981-995, 2005.

[2] H. Gunes, B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," Image and Vision Computing, vol. 31, no. 2, pp. 120-136, 2013.

[3] H. Gunes, "Automatic, dimensional and continuous emotion recognition," 2010.

[4] A. Metallinou, S. Narayanan, "Annotation and processing of continuous emotional attributes: Challenges and opportunities," 10th IEEE international conference and workshops on automatic face and gesture recognition (FG). IEEE, pp. 1-8, 2013.

[5] M. Valstar, J. Gratch, B. Schuller, et al, "AVEC 2016: Depression, mood, and emotion recognition workshop and challenge," in Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge. ACM, pp.3–10, 2016.

[6] D. Le, E. M. Provost, "Emotion recognition from spontaneous speech using hidden markov models with deep belief networks," in workshop on automatic speech recognition and understanding (ASRU). IEEE, pp. 216–221.

[7] P. Cardinal, N. Dehak, A. L. Koerich, et al, "ETS system for AVEC 2015 challenge," in Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge. ACM, pp. 17–23, 2015.

[8] F. Povolny, P. Matejka, M. Hradis, et al, “Multimodal emotion recognition for AVEC 2016 challenge,” in Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge. ACM, pp. 75–82, 2016.
[9] J. Huang, Y. Li, J. Tao, et al, “Continuous Multimodal Emotion Prediction Based on Long Short Term Memory Recurrent Neural Network,” Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge. ACM, pp. 11-18, 2017.
[10] M. A. Nicolaou, H. Gunes, M. Pantic, “Output-associative rvm regression for dimensional and continuous emotion prediction,” Image and Vision Computing, vol. 30, no. 3, pp. 186-196, 2012.
[11] M. Wöllmer, F. Eyben, S. Reiter, et al, “Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies,” Proc. 9th Interspeech 2008 incorp. 12th Australasian Int. Conf. on Speech Science and Technology SST 2008, Brisbane, Australia, pp. 597-600, 2008.
[12] T. Zhang, W. Zheng, Z. Cui, et al, “Spatial-temporal recurrent neural network for emotion recognition,” IEEE Transactions on Cybernetics, pp. 1–9, Jan. 2018.
[13] S. Khorram, Z. Aldeneh, D. Dimitriadis, et al, “Capturing Long-term Temporal Dependencies with Convolutional Networks for Continuous Emotion Recognition,” INTERSPEECH. pp.1253-1257, 2017.
[14] P. Li, Y. Song, I. McLoughlin, et al, “An Attention Pooling based Representation Learning Method for Speech Emotion Recognition,” Proc. Interspeech 2018, pp. 3087-3091, 2018.
[15] J. Huang, Y. Li, J. Tao, et al, “End-to-End Continuous Emotion Recognition from Video Using 3D Convlstm Networks,” 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 6837-6841, 2018.
[16] G. Trigeorgis, F. Ringeval, R. Brueckner, et al, “Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network,” in International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 5200–5204, 2016.
[17] Y. Tokozume, T. Harada, “Learning environmental sounds with end-to-end convolutional neural network,” 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 2721-2725, 2017.
[18] Y. Aytar, C. Vondrick, A. Torralba, “Soundnet: Learning sound representations from unlabeled video,” Advances in neural information processing systems, pp. 892-900, 2016.
[19] A. Waibel, T. Hanazawa, G. Hinton, et al, “Phoneme recognition using time-delay neural networks,” Backpropagation: Theory, Architectures and Applications, pp. 35-61, 1995.
[20] V. Peddinti, D. Povey, S. Khudanpur, “A time delay neural network architecture for efficient modeling of long temporal contexts,” Sixteenth Annual Conference of the International Speech Communication Association. 2015.
[21] H. Meng, N. Bianchi-Berthouze, Y. Deng, et al, “Time-delay neural network for continuous emotional dimension prediction from facial expression sequences,” IEEE transactions on cybernetics, vol. 46, no. 4, pp. 916-929, 2016.
[22] M. Sarma, P. Ghahremani, D. Povey, et al, “Emotion Identification from raw speech signals using DNNs,” Proc. Interspeech 2018, pp. 3097-3101, 2018.
[23] A. Vaswani, N. Shazeer, N. Parmar, et al, “Attention is all you need,” Advances in Neural Information Processing Systems. pp. 5998-6008, 2017.
[24] L. Dong, S. Xu, B. Xu, “Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition,” 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 5884-5888, 2018.
[25] F. Ringeval, B. Schuller, M. Valstar, et al, “Avec 2017: Real-life depression, and affect recognition workshop and challenge,” Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge. ACM, pp. 3-9, 2017.
[26] F. Eyben, K. R. Scherer, B. W. Schuller, et al, “The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing,” IEEE Transactions on Affective Computing, vol. 7, no. 2, pp. 190-202, 2016.
[27] F. Eyben, M. Wöllmer, B. Schuller, “Opensmile: the munich versatile and fast open-source audio feature extractor,” Proceedings of the 18th ACM international conference on Multimedia. ACM, pp. 1459-1462, 2010.
[28] B. Tadas, R. Peter, P. M. Louis, “OpenFace: An Open Source Facial Behavior Analysis Toolkit,” In: Proc. IEEE Winter Conference on Applications of Computer Vision, New York, USA, 2010.
[29] D. P. Kingma, J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.
[30] I. Lawrence, L. Kuei, "A concordance correlation coefficient to evaluate reproducibility." Biometrics, pp. 255-268, 1989.
[31] T. Dang, B. Stasak, Z. Huang, et al, “Investigating word affect features and fusion of probabilistic predictions incorporating uncertainty in AVEC 2017,” Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge. ACM, pp. 27-35, 2017.