END-TO-END CONTINUOUS EMOTION RECOGNITION FROM VIDEO USING 3D CONVLSTM NETWORKS

Jian Huang[1,3], *Ya Li,*[1] *Jianhua Tao*[1, 2,3], *Zheng Lian*[1,3], *Jiangyan Yi*[1,3]

[1]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing, China
[3]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
{jian.huang, yli, jhtao, zheng.lian, jiangyan.yi}@nlpr.ia.ac.cn

ABSTRACT

Conventional continuous emotion recognition consists of feature extraction step followed by regression step. However, the objective of the two steps is not consistent as they are parted. Besides, there is still no consensus about appropriate emotional features. In this study, we propose an end-to-end continuous emotion recognition framework which merges feature extraction and regressor into a unified system. We employ 3D convolutional networks with Long Short-Term Memory Neutral Network (ConvLSTM) to handle spatiotemporal information for continuous emotion recognition. This model is applied on AVEC 2017 database. The experiment results reveal that ConvLSTM model makes a positive effect on the performance improvement, which outperforms the baseline results for arousal of 0.583 vs 0.525 (baseline) and for valence of 0.654 vs 0.507.

Index Terms— End-to-end learning, continuous emotion recognition, 3D convolution network, ConvLSTM

1. INTRODUCTION

Automatic estimation of emotional state has a wide application in human-computer interaction [1]. Continuous emotion recognition as a function of time assigns an emotional value to every frame in a sequence. The emotional state of every frame is typically represented as a point in continuous space, such as arousal-valence space [2].

The 2017 Audio-Visual Emotion Challenge (AVEC) [3] provides a fair benchmark to evaluate various methods for non-acted spontaneous emotion recognition. The target of this challenge is the predictions of continuous arousal, valence and liking values given audio, video, and text data. Many methods have been researched for continuous emotion recognition, such as SVR [3], Convolutional Neural Networks (CNNs) [4] and Long Short-Term Memory Neutral Network (LSTM) [5].

CNNs are usually utilized to learn emotion-salient features in speech emotion recognition [6] and video emotion recognition [7]. However, 2D ConvNet only processes spatial information. When applied to video based emotion recognition, 2D ConvNet loses temporal information after every convolution operation. Therefore, 3D convolutional networks (3D ConvNet) can be utilized to capture the temporal information encoded in multiple contiguous frames, which are introduced for video action recognition [8][9]. Fan et al. [10] use 3D ConvNets to extract effective emotional features for discrete emotion recognition and won the first place in EmotiW 2016, showing the superiority of 3D features compared with 2D features [7]. 3D ConvNet has been seldom applied in continuous emotion recognition.

Compared with CNN, LSTM can learn long-term dynamic information, yield state-of-the-art results in continuous emotion recognition [11][12]. Xing et al. [13] propose ConvLSTM model to build an end-to-end trainable model for the precipitation nowcasting problem. ConvLSTM is an integrated model of CNN and LSTM [14], which has convolutional structures in both the input-to-state and state-to-state transitions. In this study, we use ConvLSTM to model emotional spatiotemporal relationships.

However, most of emotion recognition methods follow the conventional paradigm of pattern recognition, including two steps in which the first step extracts a large dimensionality of features and the second step trains a machine learning system. The discordance and variety of emotional features hampers the progress of emotion recognition. To get rid of this problem, Trigeorgis et al. [15] proposes an end-to-end model that uses a CNN to extract features from the raw speech followed by a LSTM to train continuous speech emotion recognition model. In a similar work, Bertero and Fung [16] propose a real-time CNN trained from raw speech signal for category speech emotion detection, which achieves better performance than feature-based SVM baseline.

In another study, Khorrami et al. [17] combine CNNs and Recurrent Neural Networks (RNNs) to perform continuous emotion recognition from video. A CNN is first trained for emotion regression using frame facial images. Then, the feature vector extracted from the CNN is used to train a RNN emotion recognition model. Actually, their training is a multi-stage pipeline and two networks are trained separately. Different from their work [17], we utilize a single network with ConvLSTM, whose inputs are video and outputs are

emotional predictions, to achieve end-to-end continuous emotion recognition system.

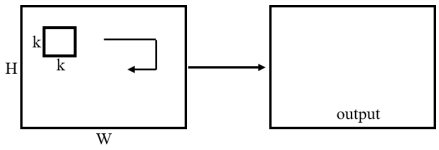
In the following, Section 2 briefly introduces the proposed model. Section 3 presents the database. Section 4 describes experiment results and analysis. Section 5 concludes this paper.

2. PROPOSED MODEL

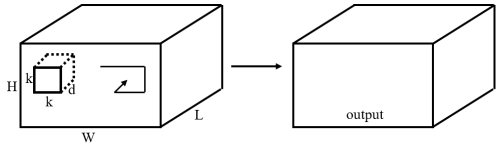
In this study, we utilize 3D convolution to achieve end-to-end manner for continuous emotion recognition. The motivation behind this idea is that, ultimately, the network learns an intermediate representation of the raw input signal automatically. The whole system only has a single network whose inputs are video and outputs are emotional predictions, as shown in Fig. 1. The system has four convolution layers including three 3D convolution layers and one ConvLSTM layer, four 3D max pooling layers followed by each convolution layer (omitted in the Fig. 1) and two inner product layers. ConvLSTM layer behind 3D convolution layers is introduced to model emotional spatiotemporal information well.

2.1. 3D ConvNet

In 2D ConvNet, convolutions focus on spatial information, thus lose temporal information of the input signal in Fig. 2(a). Whereas, 3D convolution and pooling operations have 3D kernel applied to overlapping 3D cubes spatio-temporally, and preserves the temporal information of the input signals resulting in an output volume in Fig. 2(b). Compared with 2D ConvNet, 3D ConvNet has better ability to model temporal information. We utilize 3D ConvNet to process video data directly for continuous emotion recognition.



(a) 2D convolution



(b) 3D convolution

Fig. 2. 2D convolution and 3D convolution [9]. (a). Applying 2D convolution in H×W image with k×k kernel results an image. (b) Applying 3D convolution in L×H×W video volume with d×k×k kernel results another volume. Three dimensions represent temporal, length and width respectively.

2.2. ConvLSTM

The core module of ConvLSTM can be viewed as a convolution layer embedded with a LSTM, and its convolutional layer is 3D convolution described in previous section. ConvLSTM has convolutional structures in both the input-to-state and state-to-state transitions, thus matrix multiplication is replaced by convolution operator in the formula of LSTM [13]. In addition, it determines the future state of a certain cell by the inputs and past states of its local neighbors, which enables the model to capture long-term temporal relationships. Therefore, ConvLSTM can absorb the advantage of CNN handling spatial information and LSTM handling temporal information simultaneously.

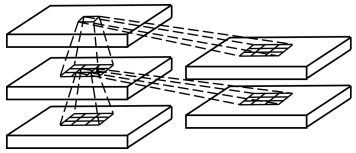


Fig. 3. Inner structure of ConvLSTM [13].

3. DATABASE

In this study, we use AVEC 2017 database based on Sentiment Analysis in the Wild (SEWA) to show the benefit of our proposed model. SEWA is a novel database of human-human interactions consisting of audio, video and text modalities. The recordings are annotated time-continuously in terms of the emotional dimensions including arousal, valence and liking. We focus on arousal and valence dimensions from video in this work. The database is divided into three subsets: Train (34 samples), Val (14 samples) and

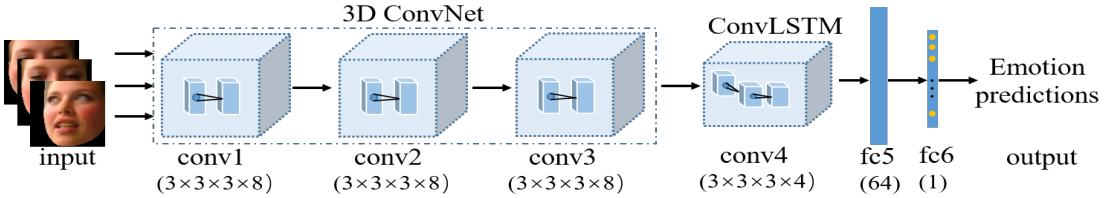


Fig. 1. Overview of proposed end-to-end continuous emotion recognition system. The architecture includes three 3D convolution layers, one ConvLSTM layer, four max pooling layers and two inner product layers.

Test (16 samples). The competition measure is the concordance correlation coefficient (CCC) [18], which combines the Pearson correlation coefficient of two times series with mean square error.

4. EXPERIMENTS AND ANALYSIS

4.1. Experiment setting

All convolution layers has 3×3×3 convolution kernels [9], are applied with padding (both spatial and temporal) and stride 1. Other parameters are shown in Fig. 1. In the beginning, four convolution layers are 3D convolution (marked as 3D ConvNet model) for comparison to proposed model (marked ConvLSTM model shown in Fig. 1). We use rmsprop optimization algorithm [19] and dropout with the rate 0.5. The maximum training epochs are 100. Our system takes full video frames as inputs, which are 112×112×1 facial images detected with OpenFace [20].

Due to no availability of test labels, we split the training set into two subsets: the first 26 subjects are used to train model and the left 8 subjects are used to adjust the parameters. The performance is evaluated on AVEC 2017 development set. We repeat the experiment five times under each parameter setting to account for instability. The final experiment results are average value of five experiments.

4.2. End-to-end recognition system with 3D ConvNet model

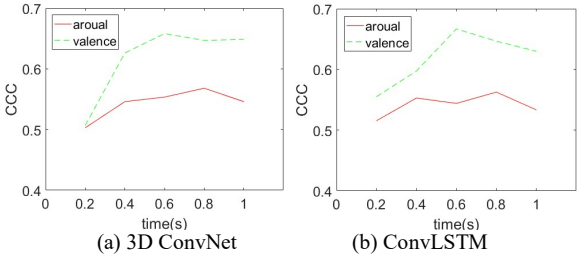
In this section, we explore the influence of max pooling and temporal pooling with 3D ConvNet model. The kernel size of four max pooling layers is shown in Table 1. The length of annotation in SEWA and the sampling rate of video are all 100 ms. Therefore, Original model ensures the length of inputs and outputs is identical. However, Original model can't fully model temporal information. Thus, Max pooling model enlarges the kernel size of temporal dimension, except the first layer with the intention of not to merge the temporal signal too early. In a result, the length of the output predictions is shortened by a factor of 6. Therefore, we repeat prediction results 6 times to compensate the reduced length.

In fact, most of samples have more than 1700 frames, demanding so much memory and training time. Besides, there also exists redundant information and label noise among adjacent frames. To relieve this problem, we utilize temporal pooling to reduce the length of the sequence data [12][21]. The temporal pooling operation adds the window to average both the features and labels, which can get the statics of the successive frames and decrease the label noise. The length of prediction results is also shortened depending on the duration time of temporal pooling. We conduct the experiments to explore appropriate duration time shown in Fig. 4(a), indicating that best duration time of arousal is 0.8s while the valence is 0.6s. Finally, we combine max pooling and temporal pooling to perform continuous emotion recognition. To be consistent with temporal pooling, the duration time of temporal pooling is 0.4s in arousal and 0.2s in valence.

Table 1: The kernel size of max pooling layers under different models.

Kernel size	Layer1	Layer2	Layer3	Layer4
Original	1×2×2	1×2×2	1×2×2	1×2×2
Max pooling	1×2×2	2×2×2	2×2×2	2×2×2
Temporal pooling	1×2×2	1×2×2	1×2×2	1×2×2
Max pooling and Temporal pooling	1×2×2	1×2×2	2×2×2	2×2×2

The experiment results in arousal and valence are shown in Table 2. Original system has worst performance. The performance of Max pooling is better than Original system, showing that gradually pooling space and time information is beneficial. The application of temporal pooling achieves best performance overall, and saves lots of training time and memory. The combination of Max pooling and Temporal pooling achieves better performance than Max-pooling, but worse than Temporal pooling. Actually, better parameters combination between max pooling and temporal pooling need to be searched for performance improvement. The above analyses apply in both arousal and valence. In addition, the performance of valence is higher than arousal.



(a) 3D ConvNet (b) ConvLSTM

Fig. 4. CCC against duration time of temporal pooling in arousal and valence.

Table 2: Performance comparisons under different models using 3D ConvNet.

CCC	Models	Arousal	valence
3D ConvNet	Original	0.429	0.524
	Max pooling	0.461	0.559
	Temporal pooling	0.558	0.656
	Max pooling and Temporal pooling	0.527	0.595

4.3. End-to-end recognition system with ConvLSTM model

It is critical to incorporate long-term temporal dependencies for continuous emotion recognition. ConvLSTM has not

only the advantages of 3D ConvNet, but also the ability of learning long-term dynamic information. Considering the scale of database and difficulties of training, we only introduce one ConvLSTM layer to model emotional spatiotemporal relationships shown in Fig. 1. The experiments similar to section 4.2 are also performed. The appropriate duration time of temporal pooling, shown in Fig. 4(b), is same as the conclusion of section 4.2.

The experiment results with ConvLSTM model are shown in Table 3. Similarly, the performance of Max pooling is better than Original system having worst results. The combination of Max pooling and Temporal pooling achieves better performance than Max-pooling but worse than Temporal pooling. Temporal pooling achieves best performance in both arousal and valence.

Table 3: Performance comparisons under different models using ConvLSTM.

CCC	Models	Arousal	Valence
ConvLSTM	Original	0.441	0.504
	Max pooling	0.456	0.532
	Temporal pooling	0.583	0.654
	Max pooling and Temporal pooling	0.546	0.624

The comparison between Table 2 and Table 3 indicates ConvLSTM achieves better performance than 3D ConvNet under all situations in arousal. The best performance of ConvLSTM is 0.583 better than 0.558 of 3D ConvNet. On the other hand, 3D ConvNet achieves better performance than ConvLSTM in valence except the last one. But 3D ConvNet and ConvLSTM achieve comparable best performance in valence, where the best performance of 3D ConvNet is 0.656 and ConvLSTM is 0.654. Therefore, ConvLSTM can improve the performance of continuous emotion recognition especially in arousal as it models the spatiotemporal relationships well.

4.4. Analysis

We take the sample *"Devel_01"* in arousal as an example to compare the effectiveness of two models, as shown in Fig. 5. 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 green and red one are the predictions of 3D ConvNet and ConvLSTM respectively. We observe that the predictions can model its dynamic trait in general, but can't reach the extremum relatively. Furthermore, the predictions of ConvLSTM get closer to the ground truth than 3D ConvNet and capture the emotional traits better.

The baseline results on development set of AVEC 2017 including video modality and multimodal, are listed in Table 4 for comparison. The results reveal that ConvLSTM achieves better performance than baseline results of video

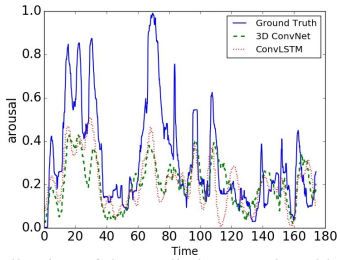


Fig. 5. A visualization of the predictions produced by 3D ConvNet and ConvLSTM against the ground truth.

modality significantly, also 0.06 higher in arousal and 0.147 higher in valence than baseline results of multimodal, which verifies the effectiveness of our proposed model. Though the performance of our models is not very excellent, we propose a compact end-to-end continuous emotion recognition system using ConvLSTM model. This work can provide a new method to improve the performance of continuous emotion recognition.

Table 4: Performance comparison between proposed method and baseline results.

CCC	Arousal	Valence
ConvLSTM	0.583	0.654
Baseline (video)	0.466	0.400
Baseline (multimodal)	0.525	0.507

5. CONCLUSION

In this paper, we use ConvLSTM model on basis of 3D convolution to build end-to-end continuous emotion system from video. The system uses a single network to merge feature extraction and regressor into a unified system. Max pooling and temporal pooling are researched to optimize recognition system. The experiment results indicate max pooling can improve the performance, but not better than temporal pooling. Temporal pooling achieves best performance and saves lots of training time and memory meanwhile. ConvLSTM achieves better performance than 3D ConvNet especially in arousal, verifying its ability of modeling emotional spatiotemporal relationships. This work provides a new method to improve the performance of continuous emotion recognition. In the future, we will explore other CNN networks to improve the performance.

ACKNOWLEDGMENTS

This work is supported by the National High-Tech Research and Development Program of China (863 Program) (No.2015AA016305), the National Natural Science Foundation of China (NSFC) (No.61425017, No. 61773379), the National Key Research & Development Plan of China (No. 2016YFB1001404) and the Major Program for the National Social Science Fund of China (13&ZD189).

6. REFERENCES

[1] Z. Zeng, M. Pantic, G.I. Roisman, et al, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," IEEE transactions on pattern analysis and machine intelligence, vol. 31, no. 1, pp.39-58, 2009.

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

[3] F. Ringeval, B. Schuller, M. Valstar, et al, "AVEC 2017Real-life Depression, and Affect Recognition Workshop and Challenge," 2017.

[4] Q. Mao, M. Dong, Z. Huang et al, "Learning salient features for speech emotion recognition using convolutional neural networks", IEEE Transactions on Multimedia, vol. 16, no. 8, pp. 2203-2213, 2014.

[5] M. Wöllmer, M. Kaiser, F. Eyben, et al, "LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework," Image and Vision Computing, vol. 31, no. 2, pp. 153-163, 2013.

[6] Q. Mao, M. Dong, Z. Huang et al, "Learning salient features for speech emotion recognition using convolutional neural networks", IEEE Transactions on Multimedia, vol. 16, no. 8, pp. 2203-2213, 2014.

[7] S.A. Bargal, E. Barsoum, C.C. Ferrer, et al, "Emotion recognition in the wild from videos using images," Proceedings of the 18th ACM International Conference on Multimodal Interaction ACM, pp. 433-436, 2016.

[8] S. Ji, W. Xu, M. Yang, et al, "3D convolutional neural networks for human action recognition," IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 1, pp. 221-231, 2013.

[9] D. Tran, L. Bourdev, R. Fergus, et al, "Learning spatiotemporal features with 3d convolutional networks," Proceedings of the IEEE international conference on computer vision, pp. 4489-4497, 2015.

[10] Y. Fan, X. Lu, D. Li, et al, "Video-based emotion recognition using cnn-rnn and c3d hybrid networks", Proceedings of the 18th ACM International Conference on Multimodal Interaction ACM, pp. 445-450, 2016.

[11] L. He, D. Jiang, L. Yang, et al, "Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks," Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge ACM, pp. 73-80, 2015.

[12] L. Chao, J. Tao, M. Yang, et al, "Long short term memory recurrent neural network based multimodal dimensional emotion recognition," Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge, ACM, pp. 65-72, 2015.

[13] S.H.I. Xingjian, Z. Chen, H. Wang, et al, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," Advances in neural information processing systems, pp. 802-810, 2015.

[14] Y. Zhao, X. Jin, X. Hu, "Recurrent convolutional neural network for speech processing," Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on IEEE, pp. 5300-5304, 2017.

[15] G. Trigeorgis, F. Ringeval, R. Brueckner, et al, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on IEEE, pp. 5200-5204, 2016.

[16] D. Bertero, P. Fung, "A first look into a Convolutional Neural Network for speech emotion detection," Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. IEEE, pp. 5115-5119, 2017.

[17] P. Khorrami, P.T. Le, K. Brady, et al, "How deep neural networks can improve emotion recognition on video data," Image Processing (ICIP), 2016 IEEE International Conference on IEEE, pp. 619-623, 2016.

[18] I. Lawrence, K. Lin, "A concordance correlation coefficient to evaluate reproducibility," Biometrics, pp. 255-268, 1989.

[19] T. Tieleman, G. Hinton, "RMSProp," COURSERA: Lecture, 2012.

[20] T. Baltrušaitis, P. Robinson, L.P. Morency, "Openface: an open source facial behavior analysis toolkit," Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on IEEE, pp. 1-10, 2016.

[21] L. Chao, J. Tao, M. Yang, et al, "Multi-scale temporal modeling for dimensional emotion recognition in video," Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, ACM, pp. 11-18, 2014.