Unified multi-stage fusion network for affective video content analysis

Yun Yi,[1,2] Hanli Wang,[2] and Pengjie Tang[2,3]
[1]*Department of Mathematics and Computer Science, Gannan Normal University, Ganzhou, P. R. China*
[2]*Department of Computer Science and Technology, Tongji University, Shanghai, P. R. China*
[3]*College of Electronics and Information Engineering, Jinggangshan University, Ji'an, P. R. China*
Email: hanliwang@tongji.edu.cn

Affective video content analysis is an active topic in the field of affective computing. In general, affective video content can be depicted by feature vectors of multiple modalities, so it is important to effectively fuse information. In this work, a novel framework is designed to fuse information from multiple stages in a unified manner. In particular, a unified fusion layer is devised to combine output tensors from multiple stages of the proposed neural network. With the unified fusion layer, a bidirectional residual recurrent fusion block is devised to model the information of each modality. Moreover, the proposed method achieves state-of-the-art performances on two challenging datasets, *i.e.*, the accuracy value on the VideoEmotion dataset is 55.8%, and the MSE values on the two domains of EIMT16 are 0.464 and 0.176 respectively. The code of UMFN is available at: https://github.com/yunyi9/UMFN.

Introduction: Affective computing has many application scenarios, with affective video content analysis being one of its active research topics [1–3]. The purpose of affective video content analysis is to automatically predict emotions elicited by videos. To this end, recent studies designed neural networks and optimized model parameters [4]. Although previous researches [5–13] have achieved promising progress, it is still challenging to analyze the emotions induced by videos.

Generally, information fusion is a significant ability of human brain, and human perceives the world by multiple modalities, *e.g.*, visual modality, aural modality, attribute modality, etc. Therefore, previous researches combine multiple features from different modalities to analyze affective video content. Moreover, previous scenes and sounds can affect the audience's current emotion, so the fusion of temporal information can improve the capability of affective video content analysis. Furthermore, the fusion of information from different stages of neural network may promote the generalization ability of network model. However, most existing works ignore information from multiple stages, or fuse them in different ways.

To address this issue, a unified multi-stage fusion network (UMFN) is proposed to uniformly fuse information from multiple stages. Specifically, a unified fusion layer (UFL) is designed to combine information from multiple stages in a unified way, and the tensor of fusion weight in UFL is learned by using an optimization algorithm. Moreover, a bidirectional residual recurrent fusion block (Bi-RRFB) is devised to model the information of each modality. Experiments are conducted on the two challenging datasets of VideoEmotion [2] and LIRIS-ACCEDE [3], and the results show that the proposed method obtains better performances than baseline methods and achieves state-of-the-art results.

In summary, the main contributions of this work are as follows. First, to predict emotions by using visual-aural features, UMFN is designed with the proposed Bi-RRFB and UFL. Second, UFL is proposed to uniformly fuse information from multiple stages, so there is no need to design different fusion layers to combine information from multiple stages. Third, Bi-RRFB is devised to model the information of each modality, in which the proposed UFL is utilized to combine the output tensors from multiple stages of Bi-RRFB.

Proposed Unified Multi-stage Fusion Network: In this section, the proposed UMFN is introduced as visualized in Fig. 1. First, feature vectors are extracted from three convolutional neural networks (CNNs). Then, the model of UMFN is trained by using feature vectors on the training set. In the testing phase, the labels of videos on the test set are predicted by utilizing the trained model.

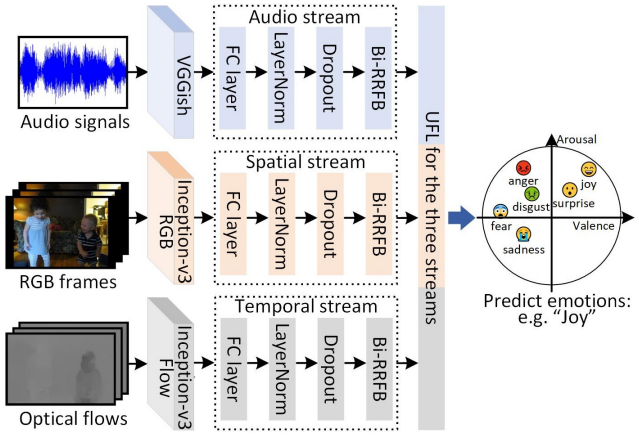

Fig 1 *Overview of the proposed UMFN.*

Network Architecture: In this work, three modalities (*i.e.*, audio signals, RGB frames and optical flows) are selected to depict videos. Similar to [12], three types of feature vectors are respectively calculated by feeding the three modalities into three CNNs, namely VGGish [14], Inception-v3-RGB [15] and Inception-v3-Flow [15].

As shown in Fig. 1, there are three streams in UMFN, *i.e.*, audio stream, spatial stream and temporal stream, which correspond to the three modalities. Except for the input dimension of the first fully connected (FC) layer of the stream, these three streams have the same network architecture. In each stream, the feature vectors of the corresponding modality are mapped to a fixed dimension by using the first FC layer of the stream. After this FC layer, layer normalization (LayerNorm) and dropout are utilized to regularize the activities of neurons. Then, the output tensors are fed into the designed Bi-RRFB, which is detailed in the next section. The output tensors from the three Bi-RRFBs are fused by using UFL, which is presented in Section 3. In all experiments, the output dimension of the first FC layer is fixed at 800, and the dropout ratio is set to 0.1.

Bidirectional Residual Recurrent Fusion Block: The proposed Bi-RRFB plays an important role in fusing information from multiple stages. In order to model temporal information, Bi-RRFB adopts the architecture of recurrent neural network (RNN). Let $\{x_t \mid t \in [1,T]\}$ and $\{h_t \mid t \in [1,T]\}$ be the input set and hidden state set of a RNN cell respectively, and T be the length of input sequence, then the output of RNN is formulated as

$$h_t = f_{rc}(x_t, h_{t-1}), \tag{1}$$

where $f_{rc}(\cdot)$ is the mapping function of the RNN cell, h_{t-1} is the hidden state at the previous time step. Although other RNN cells can be used in Bi-RRFB, the LSTM cell is selected as the basic RNN cell in Bi-RRFB.

In order to improve the generalization ability of the model, a batch of connections in the previous hidden state h_{t-1} are dropped by the dropout method, where the dropout ratio is set to 0.1. Let $X^1 = \{x_t^1 \mid t \in [1,T]\}$ and $\{h_t^1 \mid t \in [1,T]\}$ be the input set and hidden state set of the first recurrent block respectively, the output at the $t-$th temporal step is computed as

$$h_t^1 = f_{rc}(x_t^1, \text{Dropout}(h_{t-1}^1)), \tag{2}$$

where Dropout$(\cdot)$ is the dropout function. By stacking the list of h_t^1 along the temporal dimension, the output tensor at the first recurrent block H^1 can be calculated as

$$H^1 = \text{LayerNorm}(\text{Stack}(h_t^1)), \tag{3}$$

where h_t^1 can be computed by using Eq.(2), Stack$(\cdot)$ is the function that concatenates a list of tensors along the last dimension of the output tensor, and LayerNorm$(\cdot)$ is the function for layer normalization. By substituting Eq.(2) into Eq.(3), the mapping function of the recurrent block $f_{rb}(\cdot)$ can be written as

$$\begin{aligned} H^1 &= f_{rb}(X^1) \\ &= \text{LayerNorm}(\text{Stack}(h_t^1)) \\ &= \text{LayerNorm}(\text{Stack}(f_{rc}(x_t^1, \text{Dropout}(h_{t-1}^1)))). \end{aligned} \tag{4}$$

Inspired by ResNet, the input tensor of the second recurrent block X^2 is

$$\begin{aligned} X^2 &= \text{ReLU}(X^1 + H^1) \\ &= \text{ReLU}(X^1 + f_{rb}(X^1)), \end{aligned} \tag{5}$$

where $\text{ReLU}(\cdot)$ is the function of rectified linear unit. By substituting Eq.(5) into the mapping function $f_{rb}(\cdot)$, the output of the second recurrent block H^2 is calculated as

$$\begin{aligned} H^2 &= f_{rb}(X^2) \\ &= f_{rb}(\text{ReLU}(X^1 + f_{rb}(X^1))). \end{aligned} \tag{6}$$

So the output tensor of the l-th recurrent block H^l is computed as

$$\begin{aligned} H^l &= f_{rb}(X^l) \\ &= f_{rb}(\text{ReLU}(X^{l-1} + f_{rb}(X^{l-1}))), \end{aligned} \tag{7}$$

where $l \in [2, L]$ and L is the number of residual blocks.

All output tensors of recurrent blocks are fed into a multi-layer perceptron (MLP) block, which consists of two FC layers. Between the two FC layers, the methods of layer normalization and dropout are used to regularize the activities of neurons. So the output tensor of the MLP block O_m^l is calculated as

$$\begin{aligned} O_m^l &= f_{mp}(H^l) \\ &= \text{FC}(\text{Dropout}(\text{LayerNorm}(\text{FC}(H^l)))), \end{aligned} \tag{8}$$

where $f_{mp}(\cdot)$ is the mapping function of MLP block, and $\text{FC}(\cdot)$ stands for the function of fully connected layer.

In order to efficiently utilize temporal information, the proposed UFL is utilized to combine the temporal information in the output tensor O_m^l, and the output tensor O_u^l of this UFL is calculated as

$$O_u^l = f_{ul}(O_m^l), \tag{9}$$

where $f_{ul}(\cdot)$ is the mapping function of UFL and is defined as Eq.(10) in Section 3. To fully use the information of all residual blocks, UFL is employed to fuse the output tensor list $\{O_u^l \mid l \in [1, L]\}$. With regard to the combination of bidirectional RRFB, UFL is also used to combine the output tensors of the two RRFBs.

Unified Fusion Layer: In neural networks, the output information from a stage is a tensor. In order to fuse information from multiple stages uniformly, UFL is designed to combine these output tensors. In this work, the input tensors of UFL are from the outputs of 4 stages, including the temporal sequence of RNN, residual blocks, bidirectional RRFB and streams of multiple modalities. Let $X \in \mathbb{R}^{B\times K\times N}$ be the input tensor of UFL, where B is the batch size, K represents the number of categories for the classification task and is fixed at 1 for the regression task, and N stands for the number of slices of tensor X to be fused. Let $W \in \mathbb{R}^{N\times 1}$ be the weight tensor and the network parameter to be learned, the mapping function of UFL $f_{ul}(\cdot)$ is defined as

$$f_{ul}(X) = \text{Sum}(X \otimes \text{Softmax}(W)), \tag{10}$$

where $\otimes$ is the matrix multiplication with bitwise operators, $\text{Softmax}(\cdot)$ is the softmax function so that the elements of the output tensor lie in the range [0,1] and sum to 1, and the $\text{Sum}(\cdot)$ function returns the sum of each slice of the input tensor in the last dimension. Therefore, the shape of output tensor of UFL is $B \times K$.

For different tasks, N has different meanings. More specifically, N is the length of the temporal sequence of RNN in the residual block, and N is the number of residual blocks in Bi-RRFB. Furthermore, N is fixed at 2 for the fusion of output tensors of the two RRFBs. Regarding the fusion of output tensors from multiple streams, N is the number of modalities. In conclusion, tensors from multiple stages can be fused in a unified way by using UFL.

Experiment:

Dataset and Metric: Experiments are conducted on two challenging datasets, *i.e.*, VideoEmotion [2] and LIRIS-ACCEDE [3]. The VideoEmotion dataset [2] includes 1,101 user-generated videos, which fall into 8 emotional categories. For evaluation, the dataset [2] provides 10 training-test splits. In each split, the training set includes 736 videos and the test set contains 365 videos. In all experiments, we follow the official protocols, and report the mean of the 10 predicted accuracy values [2]. The MediaEval 2016 Emotional Impact of Movies Task (EIMT16) [16] is a task of the LIRIS-ACCEDE dataset [3]. EIMT16 includes 11,000 short videos, which are split into 9,800 training videos and 1,200 test videos. This dataset has two affect domains, *i.e.*, arousal and valence. According to the recommendations of the competition organizers [16], mean squared error (MSE) and Pearson correlation coefficient (PCC) are the official evaluation metrics for EIMT16, and are calculated in the two affect domains, respectively.

Experimental Setup: In this work, PyTorch is utilized to implement the proposed UMFN. Label smoothed cross entropy [17] and MSE are used as the loss function for the classification task on VideoEmotion and the regression task on EIMT16, respectively. Regarding model training, the Adam algorithm is utilized to optimize UMFN, the learning rate is fixed at 1e-5, and the parameter of weight decay is set to 1e-2 on EIMT16 and 1e-6 on VideoEmotion, respectively. To achieve a balance between performance and computational complexity, the length of input sequence T is set to 18 on EIMT16 and 50 on VideoEmotion, and the batch size for network training is configured to 32 on EIMT16 and 8 on VideoEmotion. In order to avoid over-fitting, early stopping schema is utilized. For fair comparison, UMFN uses the same scheme as the baseline methods to choose parameters.

Evaluation of Parameter L: As described in Section 3, the number of residual blocks L is an important parameter. In order to find out the best parameter, experiments are conducted on the two datasets. Table 1 shows the evaluation of parameter L, where "ACC" represents the mean of predicted accuracy values of the 10 test splits, "A-MSE" is the value of MSE in the arousal domain, and "V-MSE" stands for the value of MSE in the valence domain. For fair comparison, all methods in Table 1 utilize the same experimental setup. As shown in this table, UMFN with 3 residual blocks achieves the best results on the two datasets. Therefore, we set the parameter L to 3 in the subsequent experiments.

Table 1. Evaluation of the number of residual blocks. The best results are highlighted in bold.

Parameter L	VideoEmotion	EIMT16	
	ACC(%)	A-MSE	V-MSE
2	55.4	0.475	0.186
3	**55.8**	0.464	**0.176**
4	55.1	**0.441**	0.185
5	55.0	0.467	0.181

Evaluation of Modality Fusion: Generally, different input modalities have different representation ability. By using UMFN, experiments are performed on the two datasets to evaluate the fusion of these three modalities, *i.e.*, audio signals, RGB frames and optical flows. Table 2 reports the experimental results, where "Audio" stands for the modality of audio signals, "Flow" represents the modality of optical flows, and "RGB" is the modality of RGB frames. Except for the input modalities, all methods utilize the same experimental setup.

Table 2. Evaluation of modalities on the two datasets.

Method	VideoEmotion	EIMT16	
	ACC(%)	A-MSE	V-MSE
Audio+Flow	49.2	0.554	0.185
Flow+RGB	53.2	0.563	0.181
Audio+RGB	54.9	0.540	0.179
Audio+Flow+RGB	55.8	0.464	0.176

As shown in Table 2, the experimental results on these two datasets are improved by using UMFN to fuse the three modalities, because these modalities complement each other and UMFN reduces redundant information between these modalities. Regarding the fusion of two modal-

ities, the method "Audio+RGB" achieves better performances than the other two methods. This is probably because there is less redundancy between audio signals and RGB frames.

Comparison with Baselines: In this section, three methods are selected as baselines (*i.e.*, UMFN-GRU, UMFN-LSTM and AFRN [12]), because these methods are similar to the proposed UMFN. In particular, UMFN-GRU and UMFN-LSTM respectively use stacked GRU and stacked LSTM instead of the proposed Bi-RRFB. Moreover, AFRN utilizes LSTM to model temporal information, and combines information from two stages by using temporal-adaptive-fusion layer and multi-modal-adaptive-fusion layer. Table 3 reports the comparison between UMFN and baseline methods. To make a fair comparison, UMFN-GRU, UMFN-LSTM and UMFN utilize the same experimental setup. Regarding AFRN, Table 3 shows the results reported in [12].

Table 3. Comparison with baseline methods on the two datasets.

Method	VideoEmotion	EIMT16	
	ACC(%)	A-MSE	V-MSE
AFRN [12]	48.2	0.542	0.193
UMFN-GRU	49.2	0.847	0.186
UMFN-LSTM	50.4	1.097	0.185
UMFN	**55.8**	**0.464**	**0.176**

As shown in Table 3, AFRN obtains relatively worse results than the methods based on UMFN, because AFRN only fuses information from two stages. Moreover, UMFN-GRU and UMFN-LSTM gain similar results. In a word, the proposed UMFN obtains the best experimental results on the two datasets. This partly demonstrates that UMFN successfully fuses information from multiple stages in a unified way.

Computational Complexity: To quantify the computational time, experiments are performed using a GTX 3090 GPU. On the VideoEmotion dataset, the training time of UMFN, UMFN-LSTM and UMFN-GRU is about 1.48, 1.39 and 1.37 minutes per epoch, respectively. So, after extracting feature vectors, the model of UMFN can be trained quickly.

Comparison with State-of-The-Art Methods: This section reports the standardized evaluations between the proposed UMFN and other methods. Table 4 and Table 5 show the comparison with state-of-the-art methods on the two datasets.

Table 4. Comparison with state-of-the-art methods on VideoEmotion.

Method	ACC(%)
MMLGAN [5]	51.1
Cheng et al. [6]	52.7
KeyFrame [7]	52.9
HAMF [8]	53.1
VAANet [9]	54.5
UMFN	**55.8**

Table 5. Comparison with state-of-the-art methods on EIMT16.

Method	Arousal		Valence	
	MSE	PCC	MSE	PCC
MML [10]	1.173	0.446	0.198	0.399
Guo et al. [11]	0.543	0.459	0.209	0.326
AFRN [12]	0.542	0.522	0.193	0.468
MMLGAN [5]	1.077	0.491	0.194	0.445
AttendAffectNet [13]	0.742	0.503	0.185	0.467
UMFN	**0.464**	**0.523**	**0.176**	**0.503**

In summary, the methods [5–13] in the above tables utilize multiple modalities to describe affective content, and employ different strategies to fuse feature vectors. In particular, these methods obtain good experimental results by devising a deep visual-audio attention network, a multi-modal local-global attention network, a context-aware framework, a key frames extraction algorithm, a hierarchical attention-based multi-modal fusion network, a self-attention based network, etc. Different from the above methods, our method makes full use of the information from 4 stages, thus improving experimental performances. In conclusion, the proposed UMFN obtains better experimental results than other state-of-the-art methods on the two datasets.

Conclusion: In this work, UMFN is proposed to fuse information from multiple stages in a unified manner. For this purpose, UFL is designed to combine information uniformly, in which the weights of this layer are learned by an optimization algorithm. Meanwhile, Bi-RRFB is devised to model the information of each modality, where UFL is used to fuse information from multiple stages. On two challenging datasets, UMFN achieves better performances than baseline and other state-of-the-art methods, because information from multiple stages is fully utilized.

Acknowledgements: This work was supported in part by NSFC under Grant 61962003, 61976159, 62062041, Shanghai Engineering Research Center under Grant 17DZ2251600, and NSFJX under Grant 20202BAB202017, 20212BAB202020.

References

1. Alakus, T.B., Turkoglu, I.: Emotion recognition with deep learning using gameemo data set. Electron. Lett. 56(25), 1364–1367 (2020)
2. Jiang, Y.G., Xu, B., Xue, X.: Predicting emotions in user-generated videos. In: Proc. AAAI Conf. Artif. Intell., pp. 73–79. (2014)
3. Baveye, Y., et al.: LIRIS-ACCEDE: A video database for affective content analysis. IEEE Trans. Affect. Comput. 6(1), 43–55 (2015)
4. Sezer, A., Altan, A.: Optimization of deep learning model parameters in classification of solder paste defects. In: Proc. Int. Congr. Human-Comput. Interact., Optim. Robot. Appl., pp. 1–6. (2021)
5. Ou, Y., Chen, Z., Wu, F.: Multimodal local-global attention network for affective video content analysis. IEEE Trans. Circuits Syst. Video Technol. 31(5), 1901–1914 (2021)
6. Cheng, H., et al.: Context-aware based visual-audio feature fusion for emotion recognition. In: Proc. Int. Joint Conf. Neural Netw., pp. 1–8. (2021)
7. Wei, J., Yang, X., Dong, Y.: User-generated video emotion recognition based on key frames. Multim. Tools Appl. 80(9), 14343–14361 (2021)
8. Liu, X., Li, S., Wang, M.: Hierarchical attention-based multimodal fusion network for video emotion recognition. Comput. Intell. Neurosci. 2021, 5585041:1–5585041:11 (2021)
9. Zhao, S., et al.: An end-to-end visual-audio attention network for emotion recognition in user-generated videos. In: Proc. AAAI Conf. Artif. Intell., pp. 303–311. (2020)
10. Yi, Y., Wang, H.: Multi-modal learning for affective content analysis in movies. Multim. Tools Appl. 78(10), 13331–13350 (2019)
11. Guo, X., et al.: Global affective video content regression based on complementary audio-visual features. In: Proc. Int. Conf. Multim. Model., pp. 540–550. (2020)
12. Yi, Y., Wang, H., Li, Q.: Affective video content analysis with adaptive fusion recurrent network. IEEE Trans. Multim. 22(9), 2454–2466 (2020)
13. Thao, H.T.P., Herremans, D., Roig, G.: AttendAffectNet: Self-attention based networks for predicting affective responses from movies. In: Proc. Int. Conf. Pattern Recognit., pp. 8719–8726. (2021)
14. Hershey, S., et al.: CNN architectures for large-scale audio classification. In: Proc. Int. Conf. Acoust., Speech, Signal Process., pp. 131–135. (2017)
15. Wang, L., et al.: Temporal segment networks: Towards good practices for deep action recognition. In: Proc. Eur. Conf. Comput. Vision, pp. 20–36. (2016)
16. Dellandréa, E., et al.: The MediaEval 2016 emotional impact of movies task. In: Proc. MediaEval Workshop, (2016)
17. Szegedy, C., et al.: Rethinking the inception architecture for computer vision. In: Proc. IEEE Conf. Comput. Vision Pattern Recognit., pp. 2818–2826. (2016)

Please ensure that your Letter does not exceed the maximum length of 6 columns/3 pages.