Multi-modality Fusion for Emotion Recognition in Videos

Xinge Peng[1], Kun Li[1,*], Jiaxiu Li[1], Guoliang Chen[1] and Dan Guo[1,2,3,*]

[1]School of Computer Science and Information Engineering, School of Artificial Intelligence, Hefei University of Technology (HFUT)

[2] Key Laboratory of Knowledge Engineering with Big Data (HFUT), Ministry of Education

[3]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, China

Abstract

Video Emotion Recognition (VER) plays a crucial role in human-centered visual understanding. Current VER methods commonly leverage visual, audio and textual attribute features to train emotion classifier. Skeleton data is widely used in the action recognition field, which is used as key points to describe the movements of the whole body. Building upon this, in this paper, we introduce the integration of skeleton data into VER, aiming to learn diverse actions to improve emotion recognition performance. Specifically, we propose a model with the input of three modalities, *i.e.*, visual, audio and skeleton data. Subsequently, the NeXtVLAD module is employed to aggregate emotion clues, and the resulting features from the three modalities are concatenated to a SE block to get the rich information from three modalities. Finally, the features are fed into the classifier for emotion classification. Extensive experimental results conducted on the VideoEmotion-8 dataset demonstrate our proposed model achieves comparable performance on video emotion recognition.

Keywords

Skeleton data, action understanding, emotion recognition, video understanding

1. Introduction

Recognizing the emotional state of people in videos is a basic but challenging task in video understanding field. Many subtasks has grown out of this task, for instance, Gao *et al*. [1] proposed a new task that aimed to recognize the emotional relationship between the two interactive characters in a video. The existing method for VER task mostly utilize features of more than one modality. Specifically, Zhao *et al*. [2] proposed an end-to-end manner with visual and audio information, and Wang *et al*. [3] took the textual information as well as the visual information together. Moreover, Zhang *et al*. [4] used all of the three modalities to refine multi-modal representations and explore the commonality among different modalities in Multi-modal Multi-label Emotion Recognition task. Different from existing works, we incorporate

MiGA@IJCAI23: International IJCAI Workshop on Micro-gesture Analysis for Hidden Emotion Understanding, August 21, 2023, Macao, China.

*Corresponding authors.

xg.pengv@gmail.com (X. Peng); kunli.hfut@gmail.com (K. Li); jiaxiuli@mail.hfut.edu.cn (J. Li); guoliangchen.hfut@gmail.com (G. Chen); guodan@hfut.edu.cn (D. Guo)

0009-0006-3820-8810 (X. Peng); 0000-0001-5083-2145 (K. Li); 0009-0006-4644-5832 (J. Li); 0009-0002-7984-3184 (G. Chen); 0000-0003-2594-254X (D. Guo)

CEUR Workshop Proceedings (CEUR-WS.org)

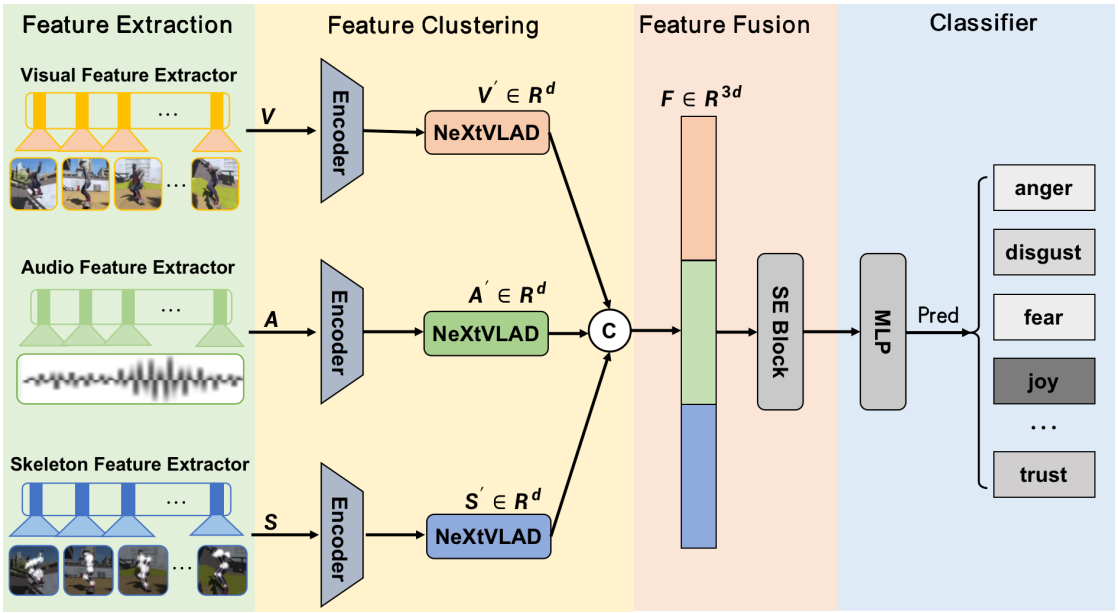

Figure 1: Overview of the proposed multi-modal emotion recognition model. Our model contains four components: (1) Feature Extraction, which is used to extract features of three modalities. (2) Feature Clustering, which include an Encoder [5] and a NeXtVLAD [6] module. (3) Feature Fusion, which consists of a SE block [7]. (4) Classifier, which generate the output class of a certain video.

for the first time the corresponding skeleton feature into the emotion recognition process. The contributions of our work are three aspects:

- We address video emotion recognition from a novel perspective by incorporating not only the conventional modalities but also the valuable skeleton information.
- We propose a multi-modality model with a visual, an audio and a skeleton stream, where each stream facilitates each modality to capture rich information from different perspectives.
- Our model achieves comparable result on VideoEmotion-8 dataset with the accuracy of 43.2.

2. Methodology

2.1. Data Preparation

As depicted in Figure 1, give the video V with the total length of T, we first extra features of three modalities. For the visual features, we utilize the Swin Transformer [8] to exact the frame-level features $V \in \mathbb{R}^{T\times 768}$. The audio features $A \in \mathbb{R}^{T\times 128}$ are exacted by the VGGish [9] model. As for the skeleton features, we utilize the BlazePose model [10] to get the pose data $S \in \mathbb{R}^{T\times 99}$. T is set to 100 in our model.

2.2. Feature Clustering and Fusion

To enhance the exacted features, we first utilize the Transformer [5] encoder layer to capture intra-modal relationships within each modality. Next, we utilize the NeXtVLAD [6] module to extract rich and distinctive information from each modality by aggregating local features and capturing global context. Finally, to refine the multi-modal representations, we combine the outputs of the three modalities using a concatenation operation, which is illustrated in Eq. 1. To sufficiently explore the diversity among different modalities and enhance the commonality of each modality, we incorporate the SE block proposed by Hu *et al.* [7]. The SE block starts with a linear layer that performs dimension reduction, resulting in the output F_1 as shown in Eq. 2. This dimension reduction step helps capture the most relevant features for each modality. Next, we apply two additional linear layers to aggregate the features from the three modalities, resulting in the generation of F_2 as formulated in Eq. 3. Finally, we incorporate F_1 and F_2 through a dot product operation and output the fused feature representation $\tilde{F}$.

$$F = [V^{'}; A^{'}; S^{'}], \tag{1}$$

$$F_1 = \delta(BN(W_1(F))), \tag{2}$$

$$F_2 = \sigma(W_3(BN(\delta(W_2F_1)))), \tag{3}$$

$$\tilde{F} = F_1 \cdot F_2, \tag{4}$$

where "[;]" is the concatenation operation. δ and σ are the ReLU and sigmoid activate function respectively.

2.3. Emotion Classification

After obtaining the enhanced feature representation $\tilde{F}$, we utilize a multi-layer perception (MLP) for the classification of emotion categories. Specifically, the initial input to the MLP is a 512-dimensional feature vector. The output of the MLP is 8-dimensional vector, corresponding to the categories of emotions being classified.

3. Experiments

3.1. Dataset

VideoEmotion-8 [11] consists of 1,101 videos collected from Youtube and Flickr with the average duration of 107 seconds. The videos are divided into eight emotion categories according to the well-known Plutchik's wheel of emotions [12], including "anger", "anticipation", "disgust", "fear", "joy", "sadness", "surprise" and "trust". These categories are popularly adopted in the existing works on emotion analysis. There are at least 100 videos per category, which is believed to have enough information to reflect the emotional relationship.

Table 1
Performance of different methods with different modalities on the VideoEmotion-8 dataset.

Method	Visaul	Audio	Skeleton	Accuracy
SentiBank [15]	✓	None	None	35.5
E-MDBM [16]	✓	✓	None	40.4
NeXtVLAD [6]	✓	✓	✓	30.2
Ours	✓	✓	✓	**43.2**

3.2. Evaluation Metric and Implementation Details

Following previous work for video emotion recognition, we report the average classification accuracy of 10 runs. In each run, the train set is made by selecting 2/3 of the data from each category, and the test set take the rest. For the encoder [5], we set the number of head in multi-head self-attention mechanism to 4. The group size in NeXtVLAD [6] module is set to 8 and the number of clusters is set to 4. In the training process, we adopt the Adam optimizer [13] with the initial learning rate 0.0001. The total epoch is set to 100 with batch-size 16. Our model is implemented in PyTorch [14].

3.3. Experimental Results

As shown in Table 1, we report the performance of our model on the VideoEmotion-8 dataset. Note that we only use a simple baseline model to verify whether the skeleton data is useful for VER. Compare with existing methods, our model achieves the comparable result with the Accuracy of 43.2. At first, we study the the effectiveness of each modality feature for VER, and the results are illustrated in Table 2. Without visual branch, we can see that the Accuracy drops to 37.4, indicating the visual information is dominant of the recognition process. As for audio branch, the result (*i.e.*, Accuracy drops from 43.2 to 39.9) shows that the audio information also plays an important role in it. The VideoEmotion-8 dataset consists of videos captured in noisy environments where multiple sounds from different sources are presented. This poses a challenge to the reliability of the audio modality compared to the other two modalities (visual and skeleton). However, our analysis shows that the inclusion of skeleton features significantly improves the performance of the video emotion recognition task. This indicates that the skeleton information, which has not been fully utilized in previous human-centered emotion recognition tasks, plays a crucial role in accurately capturing emotions in videos. Additionally, we conducted experiments to determine the optimal number of clusters (α) for the NeXtVLAD module in our model. As shown in Table 3, we found that setting α to 4 yielded the best results.

4. Conclusion

In this paper, we exploit the skeleton features for video emotion recognition task for the first time, as well as address the importance of skeleton information for this task. In addition, the proposed simple baseline achieves an accuracy 43.2 on VideoEmotion-8 dataset. We believe that this paper will inspire people to consider the skeleton data for emotion recognition tasks. In

Table 2
Ablation study results of the main component in our method on the VideoEmotion-8 dataset.

Method	Accuracy
w/o visual branch	37.4
w/o audio branch	39.9
w/o skeleton branch	38.5
Ours	**43.2**

Table 3
Ablation study results of the the number of clusters α of NeXtVLAD [6] module in our method on the VideoEmotion-8 dataset.

α	Accuracy
α=2	42.5
α=4	**43.2**
α=8	42.8
α=16	41.3

the future, we plan to exploit temporal contexts [17] to model skeleton sequence. By focusing on these aspects, we anticipate improving the accuracy of VER models.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (72188101, 62020106007, and 62272144, and U20A20183), and the Major Project of Anhui Province (202203a05020011).

References

[1] X. Gao, Y. Zhao, J. Zhang, L. Cai, Pairwise emotional relationship recognition in drama videos: Dataset and benchmark, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 3380–3389.

[2] S. Zhao, Y. Ma, Y. Gu, J. Yang, T. Xing, P. Xu, R. Hu, H. Chai, K. Keutzer, An end-to-end visual-audio attention network for emotion recognition in user-generated videos, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 303–311.

[3] Y. Wang, Y. Li, P. Bell, C. Lai, Cross-attention is not enough: Incongruity-aware multimodal sentiment analysis and emotion recognition, arXiv preprint arXiv:2305.13583 (2023).

[4] Y. Zhang, M. Chen, J. Shen, C. Wang, Tailor versatile multi-modal learning for multi-label emotion recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, 2022, pp. 9100–9108.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polo-

sukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).
[6] R. Lin, J. Xiao, J. Fan, Nextvlad: An efficient neural network to aggregate frame-level features for large-scale video classification, in: Proceedings of the European Conference on Computer Vision (ECCV) Workshops, 2018, pp. 0–0.
[7] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.
[8] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 10012–10022.
[9] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, et al., Cnn architectures for large-scale audio classification, in: 2017 ieee international conference on acoustics, speech and signal processing (icassp), IEEE, 2017, pp. 131–135.
[10] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, M. Grundmann, Blazepose: On-device real-time body pose tracking, arXiv preprint arXiv:2006.10204 (2020).
[11] Y.-G. Jiang, B. Xu, X. Xue, Predicting emotions in user-generated videos, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 28, 2014.
[12] R. Plutchik, A general psychoevolutionary theory of emotion, in: Theories of emotion, Elsevier, 1980, pp. 3–33.
[13] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, International Conference on Learning Representations (2014).
[14] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems 32, 2019, pp. 8024–8035.
[15] D. Borth, T. Chen, R. Ji, S.-F. Chang, Sentibank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content, in: Proceedings of the 21st ACM international conference on Multimedia, 2013, pp. 459–460.
[16] L. Pang, S. Zhu, C.-W. Ngo, Deep multimodal learning for affective analysis and retrieval, IEEE Transactions on Multimedia 17 (2015) 2008–2020.
[17] K. Li, D. Guo, M. Wang, Proposal-free video grounding with contextual pyramid network, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2021, pp. 1902–1910.