Joint Skeletal and Semantic Embedding Loss for Micro-gesture Classification

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

In this paper, we briefly introduce the solution of our team HFUT-VUT for the Micros-gesture Classification in the MiGA challenge at IJCAI 2023. The micro-gesture classification task aims at recognizing the action category of a given video based on the skeleton data. For this task, we propose a 3D-CNNs-based micro-gesture recognition network, which incorporates a skeletal and semantic embedding loss to improve action classification performance. Finally, we rank **1st in the Micro-gesture Classification Challenge, surpassing the second-place team in terms of Top-1 accuracy by 1.10%**.

Keywords

Micro-gesture, action classification, skeleton-based action recognition, video understanding

1. Introduction

Micro-gesture Analysis for Hidden Emotion Understanding (MiGA) is a Challenge at IJCAI 2023. It is launched based on the iMiGUE [1] and SMG [2] datasets and requires understanding emotion based on the micro-gestures (MGs). The micro-gesture classification challenge aims to recognition MGs from short video clips based on the skeleton data. The iMiGUE dataset were collected from post-match press conferences. Compared to ordinary action or gesture recognition, MGs present more challenge. MGs encompass more refined and subtle bodily movements that occur spontaneously during real-life interactions. Additionally, there is an imbalanced distribution of MGs, where 28 out of 32 categories accounted for 57.8% of the data. In this challenge, we adopt the skeleton-based recognition model PoseC3D [3] as the baseline model, and introduce semantic embedding of action label [4, 5, 6, 7] to supervise the action classification.

The main contributions of our method are summarized as follows.

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

*Corresponding authors.

kunli.hfut@gmail.com (K. Li); guodan@hfut.edu.cn (D. Guo); guoliangchen.hfut@gmail.com (G. Chen); xg.pengv@gmail.com (X. Peng); eric.mengwang@gmail.com (M. Wang)

0000-0001-5083-2145 (K. Li); 0000-0003-2594-254X (D. Guo); 0009-0002-7984-3184 (G. Chen); 0009-0006-3820-8810 (X. Peng); 0000-0002-3094-7735 (M. Wang)

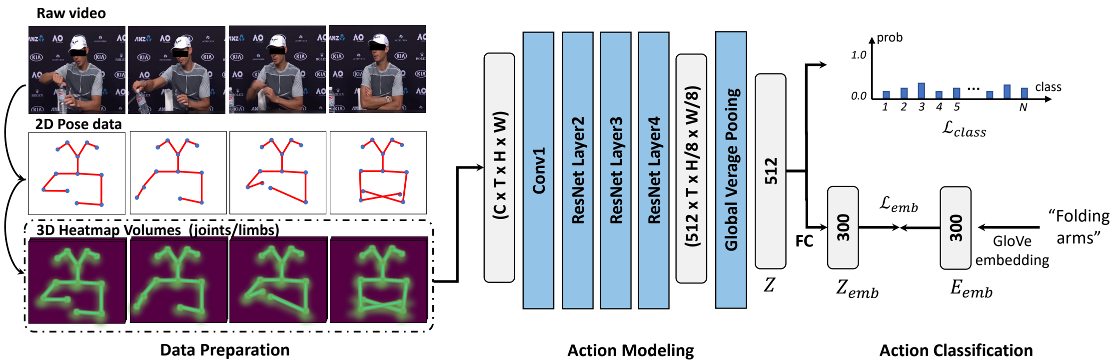

Figure 1: Overview of the proposed method for micro-gesture classification. The proposed method consists of three key steps: data preparation, action modeling, and action classification.

- We proposed a CNN-based network for micro-gesture classification. Specifically, we incorporate skeletal and semantic embedding loss for action classification.
- For the micro-gesture classification challenge, our method achieves a Top-1 accuracy of 64.12 on the iMiGUE test set. For the SMG dataset, our proposed method achieves 68.03 and 94.76 of Top-1 and Top-5 accuracy, respectively. The experimental results indicate that our method effectively captures subtle changes of micro-gestures.

2. Methodology

2.1. Data preparation

Considering that there are no lower body actions in the iMiGUE dataset, only the 22 key points extracted by OpenPose toolbox [8] of the upper body are used. For the SMG dataset, 25 keypoints of whole body are used. As shown in Figure 1, given a video $\mathbf{V}$, the extracted 2D pose data is denoted as $\mathbf{X} \in \mathbb{R}^{T \times K \times C}$, where T denotes the total frames, and K denotes the number of keypoints, and C is the number of dimension for keypoint coordinates. Then, we transform the 2D pose data $\mathbf{X}$ to a 3D heatmap volume with the size of $C \times T \times H \times W$. C is the number of joints, H and W are the height and width of the heatmap. Finally, the subjects-centered cropping and uniform sampling strategies are used to reduce the redundancy of 3D heatmap volumes. More details about the 3D heatmap volumes can refer to PoseC3D [3].

2.2. Action modeling and classification

After getting the 3D heatmap volumes, here, we use 3D-CNNs to capture the spatiotemporal dynamics of skeleton sequences. Specifically, we first use the SlowOnly [9] model as the backbone for skeleton-based action recognition. Then, we use global average pooling to generate a skeletal embedding $Z \in \mathbb{R}^{512}$. The vector Z is fed into a fully-connected (FC) layer for action classification. We also consider GloVe embedding of action label for the supervision of action classification. Specifically, we first transform the action label to 300-dimension GloVE [10] word embedding E_{emb}. Then, we use a fully-connected layer to convert the vector $Z \in \mathbb{R}^{512}$ to a 300-dimension vector $Z_{emb} \in \mathbb{R}^{300}$. Here, we use a semantic loss $\mathcal{L}_{emb}$ to make Z_{emb} close to semantic embedding E_{emb}.

2.3. Loss Optimization

$$\mathcal{L} = \mathcal{L}_{class} + \alpha \cdot \mathcal{L}_{emb}, \tag{1}$$

$$\mathcal{L}_{emb} = \|Z_{emb} - E_{emb}\|^2, \tag{2}$$

where α is a hyper-parameter to balance the two losses, and we will discuss it in the experiment. $\mathcal{L}_{emb}$ is MSE loss to supervise the semantic embedding. $\mathcal{L}_{class} = \mathcal{L}_{XE}$ is the cross-entropy loss to supervise the skeletal embedding. In addition, $\mathcal{L}_{class}$ also serve as the classification loss.

3. Experiments

3.1. Datasets

iMiGUE [1] dataset. This dataset comprises 32 MGs, along with one non-MG class, collected from post-match press conferences videos of tennis players. This challenge follows a cross-subject evaluation protocol, wherein the 72 subjects are divided into a training set consisting of 37 subjects and a testing set comprising 35 subjects. For the MG classification track, 12,893, 777, and 4,562 MG clips from iMiGUE are used for train, val, and test, respectively. **SMG [2] dataset.** This dataset consists of 3,692 samples of 17 MGs. The MG clips are annotated from 40 long video sequences, which in total contain 821,056 frames. Each long video sequence has a duration of 10-15 minutes. The dataset was collected from 40 subjects while narrating both a fake and a real story to elicit various emotional states.

3.2. Evaluation Metrics and Implementation Details

For the micro-gesture classification challenge, we calculate the Top-1 Accuracy to assess the prediction results. For the micro-gesture classification challenge, we implement our approach with the PYSKL toolbox [11]. The model is trained with SGD with momentum of 0.9, weight decay of 3e^{-4}. We set the batch size to 32, set the initial learning rate to 0.2/3. In addition, the model is trained 100 epochs with CosineAnnealing learning rate scheduler. The SlowOnly model is adopted as the 3D-CNN backbone. For the ensemble model (Joint&Limb), we use the weighted summation of scores for two modalities with a ratio of 2:3.

3.3. Experimental Results

As shown in Table 1, we report top-3 results on the test set of the iMiGUE dataset. Our team achieves the best Top-1 Accuracy of 64.12, which is higher than the runner-up by 1.10%. In addition, we also compare our approach with different skeleton-based action recognition methods on the iMiGUE and SMG datasets. At first, we investigate the impact of hyper-parameter α in Eq. 1. As shown in Table 2, the proposed method achieves the best Top-1 accuracy when $\alpha = 20$. Thus, we set $\alpha = 20$ as the optimal setting in the following experiments. Secondly, as shown in Table 3, on the iMiGUE dataset, our method achieves the best Top-1 and

Table 1
The top-3 results of micro-gesture classification on the iMiGUE test set. Data is provided by the Codalab competition page[1].

Rank	Team	Top-1 Accuracy (%)
1	**Ours**	**64.12**
2	NPU-Stanford	63.02
3	ChenxiCui	62.63

Table 2
Ablation study results of α on the iMiGUE test set with joint features.

Parameter	Top-1 (%)	Top-5 (%)
α=1	59.58	90.05
α=10	60.37	90.03
α=20	**62.28**	**90.62**
α=30	61.03	89.89
α=40	60.21	90.11
α=50	61.60	90.31

Table 3
The results of micro-gesture classification on the iMiGUE and the SMG test sets.

Method	Modality	iMiGUE dataset		SMG dataset	
		Top-1 (%)	Top-5 (%)	Top-1 (%)	Top-5 (%)
ST-GCN [12]	Joint	46.38	85.47	58.03	93.61
ST-GCN++ [11]	Joint	49.56	85.09	58.03	93.61
StrongAug [11]	Joint	53.13	87.00	62.79	92.62
AAGCN [13]	Joint	54.73	84.59	60.49	91.64
CTR-GCN [14]	Joint	53.02	86.19	60.98	90.82
DG-STGCN [15]	Joint	49.56	85.09	65.57	90.82
PoseC3D [3]	Joint	59.54	89.59	63.44	88.20
PoseC3D [3]	Limb	60.74	90.51	63.11	93.77
Ours	Joint	62.28	90.62	66.07	91.80
Ours	Limb	63.48	91.01	65.57	92.62
Ours	Joint&Limb	**64.12**	**91.10**	**68.03**	**94.76**

Top-5 accuracy of 64.12 and 91.10, respectively. Compared with the baseline model PoseC3D, our method exhibits 2.74% improvements in Top-1 accuracy with the joint feature as input. On the SMG dataset, our method also achieves the best performance (*i.e.*, 68.03 and 94.76 of Top-1 and Top-5 accuracy). Compared with the PoseC3D model, our method achieves 2.63% improvements in Top-1 accuracy in terms of joint feature. In addition, we can see that the ensemble model (Joint&Limb) also shows significant performance improvement (*i.e.*, 1.96% and 2.96% improvements on Top-1 and Top-5 accuracy compared with 'Joint').

[1]The Codalab competition page: link

4. Conclusions

In this paper, we present our solution developed for the MiGA challenge hosted at IJCAI 2023. Our approach adopts the PoseC3D model as a baseline, incorporating both skeletal embedding loss and semantic embedding loss. By leveraging the joint and limb modality data, our approach achieved the first place with the top-1 and top-5 accuracy of 64.12 and 91.10, respectively. In the future, we plan to address the issues in this challenge from other perspectives, *e.g.*, more robust network for human pose estimation, data augmentation for imbalanced data learning, RGB-based visual feature for micro-gesture recognition, and temporal context modeling [16, 17] for capturing subtle changes of MG.

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. Liu, H. Shi, H. Chen, Z. Yu, X. Li, G. Zhao, imigue: An identity-free video dataset for micro-gesture understanding and emotion analysis, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 10631–10642.

[2] H. Chen, H. Shi, X. Liu, X. Li, G. Zhao, Smg: A micro-gesture dataset towards spontaneous body gestures for emotional stress state analysis, International Journal of Computer Vision 131 (2023) 1346–1366.

[3] H. Duan, Y. Zhao, K. Chen, D. Lin, B. Dai, Revisiting skeleton-based action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 2969–2978.

[4] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, T. Mikolov, Devise: A deep visual-semantic embedding model, Advances in Neural Information Processing Systems 26 (2013).

[5] M.-C. Yeh, Y.-N. Li, Multilabel deep visual-semantic embedding, IEEE Transactions on Pattern Analysis and Machine Intelligence 42 (2019) 1530–1536.

[6] Z. Wei, J. Zhang, Z. Lin, J.-Y. Lee, N. Balasubramanian, M. Hoai, D. Samaras, Learning visual emotion representations from web data, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 13106–13115.

[7] P. P. Filntisis, N. Efthymiou, G. Potamianos, P. Maragos, Emotion understanding in videos through body, context, and visual-semantic embedding loss, in: Proceedings of the ECCV 2020 Workshops, 2020, pp. 747–755.

[8] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, Y. A. Sheikh, Openpose: Realtime multi-person 2d pose estimation using part affinity fields, IEEE Transactions on Pattern Analysis and Machine Intelligence (2019).

[9] C. Feichtenhofer, H. Fan, J. Malik, K. He, Slowfast networks for video recognition, in:

Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6202–6211.
[10] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014, pp. 1532–1543.
[11] H. Duan, J. Wang, K. Chen, D. Lin, Pyskl: Towards good practices for skeleton action recognition, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 7351–7354.
[12] S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 32, 2018.
[13] L. Shi, Y. Zhang, J. Cheng, H. Lu, Skeleton-based action recognition with multi-stream adaptive graph convolutional networks, IEEE Transactions on Image Processing 29 (2020) 9532–9545.
[14] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, W. Hu, Channel-wise topology refinement graph convolution for skeleton-based action recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 13359–13368.
[15] H. Duan, J. Wang, K. Chen, D. Lin, Dg-stgcn: Dynamic spatial-temporal modeling for skeleton-based action recognition, arXiv preprint arXiv:2210.05895 (2022).
[16] D. Guo, S. Wang, Q. Tian, M. Wang, Dense temporal convolution network for sign language translation, in: Proceedings of the International Joint Conferences on Artificial Intelligence, 2019, pp. 744–750.
[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.