A Multimodal Micro-gesture Classification Model Based on CLIP*

Yiwen Wang[1], Zhenyang Dong[1], Pengxia Li[1] and Yujie Liu[1,*]

[1]College of Computer Science and Technology, Qingdao Software College, China University of Petroleum, Qingdao 266580

Abstract

This paper primarily introduces our approach in the 2nd MiGA-IJCAI Challenge Track 1, which focuses on micro-gesture recognition. The micro-gesture dataset has the characteristics of small action amplitude, short duration, and concentrated actions in specific parts. Regarding these issues, We propose a multimodal micro-gesture recognition network based on CLIP. In the video modality, we use a frozen CLIP model as the teacher network and train the student model via distillation. For the skeleton modality, we convert the data into 3D heatmaps, reducing the inherent sparsity of skeleton data. Additionally, we apply text features learned from CLIP to the skeleton modality, enabling interaction between the two models. Our approach achieved an accuracy of 68.9% in micro-gesture recognition.

Keywords

Micro-gesture, action classification, CLIP, Vision-Language Model

1. Introduction

Pose recognition refers to the automatic recognition and analysis of human posture and movements through computer technology. It can involve identifying information such as human posture, actions, and posture angles to infer the state and intention of the human body. Micro-gesture classification is a critical research direction in the field of computer vision. In this domain, most efforts are dedicated to recognizing descriptive gestures. "`Descriptive gestures`" refer to purposeful and more prominent body movements, such as drinking water or running, through which people can clearly express their emotions. However, in certain contexts like interviews and competitions, individuals may deliberately hide their true feelings, making it difficult for computers to further analyze their emotions. In contrast, "`Micro-gestures`" are spontaneous, unconscious subtle movements that can provide valuable insights into an individual's internal state, revealing hidden emotional conditions. This makes micro-gesture detection significant in psychology, behavior analysis, and communication studies.

Gesture recognition typically relies on video or skeleton data. Video data usually contains richer information but requires more computational resources and time to process. Skeleton data can provide abstract gesture information, reducing the impact of background noise, but its

MiGA@IJCAI24: International IJCAI Workshop on 2nd Micro-gesture Analysis for Hidden Emotion Understanding, August 3, 2024, Jeju, South Korea.

*Corresponding author.

†These authors contributed equally.

z23070063@upc.edu.cn (Y. Wang); 1907040223@s.upc.edu.cn (Z. Dong); s23070018@s.upc.edu.cn (P. Li); liuyujie@upc.edu.cn (Y. Liu)

0009-0006-2788-5445 (Y. Wang)

sparsity may lead to the loss of some detailed information. Therefore, there are currently many works for multi-modal recognition. DCSNet[1]utilizes the complementary information between RGB and skeleton modes, and uses the human skeleton as guidance information to crop out key activity areas of the human body in RGB frames for recognition, greatly eliminating background interference. VPN [2] generates feature maps that are more discriminative for subtle actions through spatial embedding and attention networks. S Kim [3] proposed a Transformer model based on 3D deformable attention, which can better learn spatiotemporal attention for cross membrane action recognition. However, these methods are all models trained from scratch and have high computational complexity.

In the video modality, Transformer-based methods [4, 5] dominate due to their ability to capture long-range temporal dependencies, which better understand temporal action sequences in videos. However, Transformer models usually require large datasets to fully utilize their powerful parameterization capabilities. With the advent of CLIP [6] in recent years, the large-scale pre-training on image-text pairs has addressed some limitations of Transformers, allowing for more effective use of large-scale data and enabling transfer learning across various tasks. This has also led to improved performance in gesture recognition tasks [7, 8]. In this paper, we consider two aspects:

1. Micro-gesture recognition needs more attention to detail, and background information should not be overly emphasized;
2. Training video and skeleton multimodal models is computationally expensive.

Therefore, In order to reduce the high computational complexity caused by using multimodal data, based on Froster CLIP [9], we propose a token attenuation strategy in the video encoding module, we delete a portion of tokens based on attention weights every time we pass through the Transformer, gradually filtering out unimportant tokens layer by layer. Experimental results demonstrate the effectiveness of this method.

For the skeleton modality, to align with the video modality's CLIP model, we apply the text embeddings learned from the video modality to the skeleton network. The PoseConv-3D model is specifically augmented with CLIP text embeddings [10], facilitating collaboration with the CLIP text encoder. With the integration of CLIP text embeddings, the model is enabled to work collaboratively. Through the comprehensive utilization of feature extraction methods from both skeleton and video modalities, the performance of micro-gesture recognition tasks can be enhanced by leveraging both skeleton sequences and video images.

Our method's main contributions are as follows:

- In the video modality, we propose a video action recognition network based on CLIP. Specifically, we enhance the focus on details by implementing a token weight attenuation strategy.
- In the skeleton modality, we apply the CLIP text embeddings trained in the video model to a 3D-CNN network, improving the correlation between the models.
- In the micro-gesture classification competition, our method achieved an accuracy of 68.9% on the IMIGUE dataset. Experimental results demonstrate that this approach effectively recognizes micro-gestures.

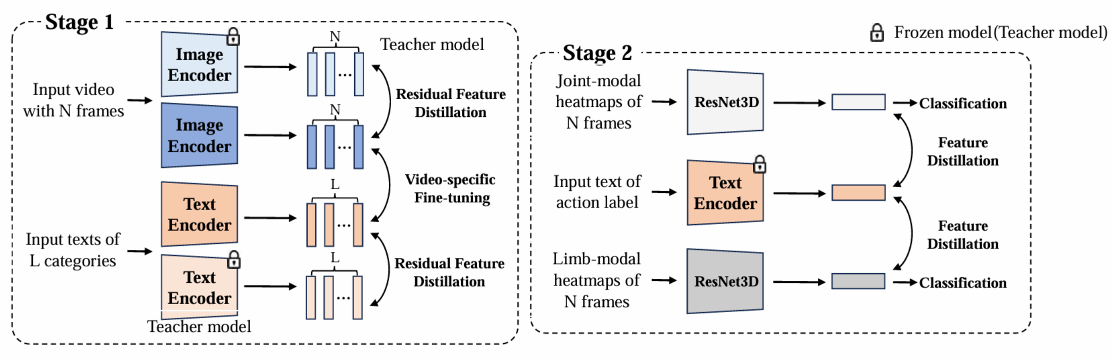
Stage 1
Input video with N frames
Image Encoder
Image Encoder
N
N
Teacher model
Residual Feature Distillation
Video-specific Fine-tuning
Input texts of L categories
Text Encoder
Text Encoder
L
L
Residual Feature Distillation
Teacher model
Frozen model(Teacher model)
Stage 2
Joint-modal heatmaps of N frames
ResNet3D
Classification
Feature Distillation
Input text of action label
Text Encoder
Feature Distillation
Limb-modal heatmaps of N frames
ResNet3D
Classification

Figure 1: Illustration of proposed training method. Our practices follow the pipeline: (left) using the frozen CLIP to distill knowledge at a student VCLIP model and (right) training ResNet3D to extract the features of heatmaps with the help of distilled text encoder.

2. Methodology

2.1. Data Preprocessing

For video data, the first step is to segment the entire video into smaller clips containing actions. Each video clip can be represented as $x^v \in \mathbb{R}^{T\times H\times W\times 3}$, where H,W indi cate resolution, T represents the number of frames.

For skeleton data, the input is represented as $X \in \mathbb{R}^{C\times N\times T}$, where $C = 3$ represents the coordinate dimensions, $N = 22$ represents the number of keypoints, and T represents the number of frames. Then, the skeleton data is represented as a heatmap of size $N \times T \times H \times W$, where H and W represent the height and width of the image, respectively. Each heatmap, centered around a keypoint, is composed of K Gaussian heatmaps to obtain the heatmap J, where K represents the number of joint points.

2.2. Video-specific Fine-tuning with Distillation

As shown in Figure 1 is our model structure. To apply CLIP for video action recognition task, FROSTER[9] introduces distillation into their full fine-tuning method, demonstrating superior performance. Given a video clip x and textual prompts of all categories, they are processed by frozen CLIP's vision encoder and text encoder respectively to obtain frame-specific visual features z_f^v and textual features z_f^t, denoted as $z_f^v \in \mathbb{R}^{T\times C}$ and $z_f^t \in \mathbb{R}^{L\times C}$. Here, C means the dimension of extracted features, L denotes the number of classes and T denotes the number of frames. Similarity, an improved VCLIP student model (see Section 3.3) converts the visual and textual data to corresponding embeddings z_g^v and z_g^t, whose shape is the same as the frozen branch counterpart's. In most cases, fully fine-tuning method on CLIP directly calculate the similarity between the z_g^v and z_g^t and use a cross-entropy loss function to optimize the tuned model, which is defined as:

$$\ell_{cls} = -\frac{1}{N}\left[\sum_{i=1}^{N} \ell_{ce}\{sim[p(z_g^{v,i}), z_g^t]/\tau, y_i\}\right], \tag{1}$$

where $y_i \in \mathbb{R}^L$ represents the ground truth, $p(\cdot)$ denotes temporal average pooling strategy, $sim(\cdot,\cdot)$ denotes cosine similarity calculation and τ is a temperature parameter. However, Froster attempts to enhance model's generalization ability by additionally introducing a residual MLP structure and distillation method. Specifically, the tuned features are transformed as follow:

$$\hat{z}_g = z_g + \alpha \times \mathrm{MLP}(z_g), \tag{2}$$

where α is a balancing coefficient. For simplicity, z_g^v and z_g^t are uniformly represented as z_g, and similar simplifications have been made in next formula. Then the distillation process can be written as:

$$\ell_{fd} = \sum_i^N \|(z_f - \hat{z}_g)\|_2. \tag{3}$$

The overall loss function is defined as:

$$\mathscr{L} = \ell_{cls} + \beta(\ell_{fd}^v + \ell_{fd}^t), \tag{4}$$

where β is a balancing coefficient.

2.3. Improved VCLIP

When tuning a pretrained CLIP into video downstream task, one question is how to capture temporal relationships in videos.[11] solve this problem by expanding the temporal attention view. Specifically, normal self-attention mechanism proposed in [12]operates as follows:

$$\begin{aligned} S(Q,K) &= \mathrm{softmax}(\frac{QK^T}{\sqrt{d}}), \\ Y &= S(Q,K)V, \end{aligned} \tag{5}$$

where $S(Q,K) \in \mathbb{R}^{N\times N}$ means the similarity matrix, $Q \in \mathbb{R}^{N\times d}$, $K \in \mathbb{R}^{N\times d}$ and $V \in \mathbb{R}^{N\times d}$ means query, key and value features, T means transpose operation, d refers to the dimension of Q and N is the number of tokens. It is clear that self-attention fails to boost the interaction within inter-frame information in this case. So VCLIP aggregates the temporal information by concatenating K_p, K_c and K_f along token dimension, while K_p, K_c and K_f represent original key features of the previous, current and following frames respectively. Similarly, V is converted to its temporal version as well. In this way, VCLIP can model spatial-temporal correlation jointly without extra parameters. However, we notice that the growing number of tokens greatly increases unnecessary compute costs, since redundant information in consecutive frames is also aggregated. To overcome this issue, we design a tokens-decay strategy based on vector similarity. Given the query features of [cls] token denoted as q_{cls}, we remain k tokens in Y and drop the others according to the top-k scores in similarity vector $S(q_{cls}, K_c) \in \mathbb{R}^T$. Note [cls] token is a fixed reserved token and will not be involved in the filtering process.

2.4. Action Modeling and Classification with Heatmaps

Assuming we have obtained the heatmaps of any modal(joint/limb). Then, following the practices in [10], we extract clip-level features via 3D ResNet-50 network and map these features into the same dimension as z_g^t (see Section 3.3). For simplicity, we use $\hat{x}$ represent the mapped embedding. We employ the text encoder trained in Section 3.2 to generate well aligned embedding as an auxiliary supervision. Similar to common classification tasks, $\hat{x}$ is fed into a linear layer to classify. Finally, the overall loss function used to optimize the 3D temporal network is defined as:

$$\begin{aligned} \ell_{cls} &= -\frac{1}{N}\sum_{i}^{N}\sum_{j}^{K} y_{i,j}\log \hat{y}_{i,j}, \\ \ell_{emb} &= -\frac{1}{N}\sum_{i}^{N}\sum_{j}^{K} y_{i,j}\|\hat{x}_i - z_g^{t,i}\|^2, \\ \mathscr{L} &= \ell_{cls} + \gamma \cdot \ell_{emb}, \end{aligned} \tag{6}$$

where γ is a balancing coefficient. We separately train two models for each modality and fuse their classification results in Section 3.5.

2.5. Ensemble

Along with the similarity-based logits from Froster, we have two sets of head-based logits derived from Resnet3D, each set corresponding to a unique modality(joint/limb). We used a fixed-step weight search strategy to find an optimal fusion weight. The result can be formulated as follows:

$$Y = w_1 * y_1 + w_2 * y_2 + w_3 * y_3. \tag{7}$$

3. Experiments

3.1. Datasets

iMiGUE[13] dataset. The iMiGUE dataset is derived from post-match interview videos with athletes. After an intense competition, a professional athlete needs to be interviewed by reporters. In these videos, a total of 18,499 micro-gesture samples were annotated, divided into 32 category labels. On average, each video contains about 51 micro-gesture samples. The duration of these micro-gesture samples varies from 0.18 seconds to 80.92 seconds, with an average duration of 2.55 seconds.

3.2. Comparison to State-of-the-art Methods

We validated our proposed method through comparative analysis on the IMIGUE dataset. As shown in Table 1, we compared our approach with state-of-the-art methods, including the GCN-based Hyperformer method for the skeleton modality, the PoseC3D method based on 3D-CNN, and the Frozen CLIP method for the video modality. Compared to the DSCNet method, which also employs CNNs in a multimodal approach, our method improved the Top-1 accuracy by 6.37%. Moreover, compared to the Frozen CLIP method that uses a text encoder, our method significantly enhanced the Top-1 accuracy by 11.11%, demonstrating the effectiveness of our approach.

Table 1
Comparison with Other Methods on the iMiGUE Dataset

Methods	Modality	Top-1(%)	Top-5(%)
ST-GCN [14]	Skeleton	46.38	85.47
Hyperformer [15]	Skeleton	57.01	87.86
PoseC3D [10]	Skeleton+Joint	59.54	89.59
PoseC3D [10]	Skeleton+Limb	60.74	90.51
TRN [16]	RGB	55.24	89.17
Frozen clip [7]	RGB	57.79	91.89
Froster clip [9]	RGB	61.05	90.03
DSCNet [1]	RGB+Skeleton	62.53	92.41
Ours	RGB+Skeleton	68.90	92.43

Table 2
The Impact of Different Weights on the Results

RGB	Skeleton(Joint)	Skeleton(Limb)	Top-1(%)
1.0	0.0	0.0	61.31
0.0	1.0	0.0	63.83
0.0	0.0	1.0	54.65
0.2	0.4	0.4	65.47
0.3	0.35	0.35	66.92
0.4	0.3	0.3	67.90
0.55	0.4	0.05	68.90
0.5	0.25	0.25	68.30
0.6	0.2	0.6	67.93

3.3. Ablation Study

We investigated the impact of different weights on the results, assigning different weights to three models. As shown in Table 2, the results revealed that the highest accuracy was achieved when the weights for RGB, Joint, and Limb models were set to 0.55, 0.4, and 0.05 respectively. The accuracy on the IMIGUE dataset reached 68.90% under these weight configurations.

We explored the impact of the Attenuation token on the results. As shown in Table 3, the Attenuation token strategy was adopted in the Transformer module, where tokens were attenuated based on attention weights, filtering out unimportant tokens to make the model focus more on crucial parts, thus enhancing the recognition accuracy. Additionally, reducing the number of tokens can decrease computational complexity. On the IMIGUE test set, we improved the accuracy by 0.26% while reducing memory usage to 1.28G, thus enhancing the overall performance.

Table 3
The Impact of Attenuation Token on the Results

Model	Attenuation token	Top-1(%)	Mem(G)
Froster clip	×	61.05	1.88
Ours	√	61.31	1.28

4. Conclusions

In this paper, we presented our solution for the MIGA Challenge organized by IJCAI 2024. Our approach involves a multimodal model based on CLIP. In the RGB modality, we proposed an attenuation token strategy building upon Froster CLIP as the baseline. In the skeleton modality, we integrated the text encoder from CLIP into PoseC3D to enhance interaction between the two modalities. Ultimately, our multimodal approach achieved an accuracy of 68.9%. In the future, we plan to address the strengths and weaknesses of both video and skeleton modalities by implementing targeted complementary operations. For example, leveraging the sparsity of skeletons to crop videos could improve the capture of detailed information in micro-gestures.

References

[1] Q. Cheng, J. Cheng, Z. Liu, Z. Ren, J. Liu, A dense-sparse complementary network for human action recognition based on rgb and skeleton modalities, Expert Systems with Applications 244 (2024) 123061. URL: https://www.sciencedirect.com/science/article/pii/S0957417423035637. doi:https://doi.org/10.1016/j.eswa.2023.123061.

[2] S. Das, S. Sharma, R. Dai, F. Bremond, M. Thonnat, Vpn: Learning video-pose embedding for activities of daily living, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16, Springer, 2020, pp. 72–90.

[3] S. Kim, D. Ahn, B. C. Ko, Cross-modal learning with 3d deformable attention for action recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 10265–10275.

[4] J. Yang, X. Dong, L. Liu, C. Zhang, J. Shen, D. Yu, Recurring the transformer for video action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 14063–14073.

[5] D. Ahn, S. Kim, H. Hong, B. C. Ko, Star-transformer: A spatio-temporal cross attention transformer for human action recognition, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023, pp. 3330–3339.

[6] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, Sastry, Learning transferable visual models from natural language supervision, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 8748–8763.

[7] Z. Lin, S. Geng, R. Zhang, P. Gao, Frozen clip models are efficient video learners, in: Computer Vision – ECCV 2022, Springer Nature Switzerland, Cham, 2022, pp. 388–404.

[8] B. Ni, H. Peng, M. Chen, S. Zhang, Expanding language-image pretrained models for

general video recognition, in: S. Avidan, G. Brostow, M. Cissé (Eds.), Computer Vision – ECCV 2022, Springer Nature Switzerland, Cham, 2022, pp. 1–18.

[9] X. Huang, H. Zhou, K. Yao, K. Han, Froster: Frozen clip is a strong teacher for open-vocabulary action recognition, 2024. arXiv:2402.03241.

[10] H. Duan, Y. Zhao, K. Chen, D. Lin, B. Dai, Revisiting skeleton-based action recognition, 2022. arXiv:2104.13586.

[11] Z. Weng, X. Yang, A. Li, Z. Wu, Y.-G. Jiang, Open-vclip: Transforming clip to an open-vocabulary video model via interpolated weight optimization, 2023. arXiv:2302.00624.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2023. arXiv:1706.03762.

[13] X. Liu, H. Shi, H. Chen, Z. Yu, X. Li, G. Zhaoz, imigue: An identity-free video dataset for micro-gesture understanding and emotion analysis, 2021. arXiv:2107.00285.

[14] S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, Proceedings of the AAAI Conference on Artificial Intelligence 32 (2018). URL: https://ojs.aaai.org/index.php/AAAI/article/view/12328. doi:10.1609/aaai.v32i1.12328.

[15] Y. Zhou, Z.-Q. Cheng, C. Li, Y. Fang, Y. Geng, X. Xie, M. Keuper, Hypergraph transformer for skeleton-based action recognition, 2023. arXiv:2211.09590.

[16] B. Zhou, A. Andonian, A. Oliva, A. Torralba, Temporal relational reasoning in videos, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018.