Micro-gesture Online Recognition with Dual-stream Multi-scale Transformer in Long Videos

Yuhan Wang[1], KeRui Linghu[1], Hexiang Huang[1] and Zhaoqiang Xia[1,2,*]

[1]School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710129, China
[2]Innovation Center NPU Chongqing, Northwestern Polytechnical University, Chongqing 400000, China

Abstract

Micro-gestures are increasingly recognized as a key indicator in the field of emotion analysis and have garnered growing interest within the field. The majority of research efforts have been directed towards the classification of micro-gestures, which entails predicting their categories. However, comparatively fewer studies have been dedicated to the detection of micro-gestures. Micro-gesture online recognition (spotting), which involves predicting both the temporal position and the category, is a preliminary step for classification but has received limited attention. In this context, we construct a deep network with dual-stream input for micro-gesture online recognition. Specifically, we utilize a sequential action recognition model to extract motion features from RGB and skeleton sequences separately, which are then processed by the multi-scale Transformer encoder as detection model. The proposed network are trained in a two-stage strategy and combined to perform the temporal spotting. Our proposed method is validated on the SMG dataset and has achieved the **first ranking** in the task of online recognition from the MiGA2024 Challenge Track 2.

Keywords

Micro-gesture online recognition, Dual-stream network, Multi-scale Transformer,

1. Introduction

In everyday interactions, humans often rely on physical gestures to express and perceive emotions, which plays a crucial role in facilitating communication and understanding among individuals. With the growing demand for intelligent systems such as robots and other human-computer interaction systems, the ability to recognize and respond to the user's emotions according to the user's body gestures has become a key component [1]. Micro-gesture is a kind of spontaneous, unconscious body gesture that is more subtle and involuntary than ordinary body gestures such as waving hands [2], and when people attempt to conceal their real feelings, micro-gestures can happen as biting finger and folding arms. As this kind of gestures is typically performed unconsciously and unintentionally, they can reveal the hidden emotional status of human beings, which is the emotional status that people express intentionally. Psychological studies [3] also show that MiGs can be more reliable as emotion indicators. So micro-gestures deserve further study in the field of emotion analysis.

MiGA2@IJCAI24: International IJCAI Workshop on 2nd Micro-gesture Analysis for Hidden Emotion Understanding, August 3, 2024, Jeju, Korea.

[*]Corresponding author.

wwww@mail.nwpu.edu.cn (Y. Wang); linghu@mail.nwpu.edu.cn (K. Linghu); huanghexiang@mail.nwpu.edu.cn (H. Huang); zxia@nwpu.edu.cn (Z. Xia)
0000-0003-0630-3339 (Z. Xia)

In recent years, the application of computer vision techniques to micro-gesture analysis has motivated significant interest. The studies in this filed can be broadly categorized into two primary classes [4]: the categorization of body gestures and the identification and localization of temporal body gestures within long sequences, also known as online recognition or spotting. Scholars have predominantly focused on the former, engaging in the classification of pre-segmented clips, where state-of-the-art methodologies have achieved notably promising performance [5, 6]. In contrast, the latter task, which entails the detection and recognition of micro-gestures within a continuous sequence, is currently short of automated methods. This gap underscores the imperative to develop an automated model for micro-gesture temporal detection, which would facilitate more precise and efficient analysis of micro-gestures, and is a critical component for the accurate interpretation and understanding of human emotions. To bridge this gap, Chen et al. [1] adopt a Deep Belief Network (DBN) with a hidden Markov model (HMM) to detect and recognize the micro-gestures in given sequences. Besides this work, they also propose an attention-based Long Short-Term Memory (BiLSTM) network to model the local temporal dynamics in micro-gesture frames and use an HMM model to enhance inference reasoning [4]. Guo et al. [7] utilize a graph-convolution based Transformer module to extract motion features of 2D skeleton sequences, and then feed them into a multi-scale Transformer for detecting the micro-gestures.

In this paper, we propose a deep network for detecting micro-gestures to locate and recognize micro-gestures from long video sequence with dual-stream input. We choose to utilize a 3D convolution network to extract multi-view motion features from RGB and 3D heatmaps converted from 2D skeleton joints. Then the multi-scale Transformer encoders are used to obtain regression and classification results of the dual input streams respectively which are fused later to improve the spotting performance. Concretely, the multi-scale Transformer encoder contains a feature pyramid module to get hierarchical multi-scale features and a local Transformer module to model the similarity between micro-gesture frames. The proposed baseline are trained in a two-stage strategy and combined to perform the temporal spotting. The main contributions of this paper can be summarized as:

- We design a deep network for dual-stream MiG online recognition based on multi-scale Transformer.
- We employ a 3D convolution network as a feature extractor and multi-scale Transfomer encoder as the detection model, which combine the feature pyramid and local Transformer to locate micro-gestures, which are trained separately in a two-stage way.
- We achieve the first ranking in the Track 2 of MiGA2024 challenge for the task of online recognition.

2. Methodology

2.1. Overall Architecture

In order to spot the micro-gestures using RGB and skeleton data, our proposed network mainly consists of two important components: 3D convolutional network (i.e., RGBPose-Conv3D [8]) and multi-scale Transformer encoder [9], the latter can be divided into three critical parts:

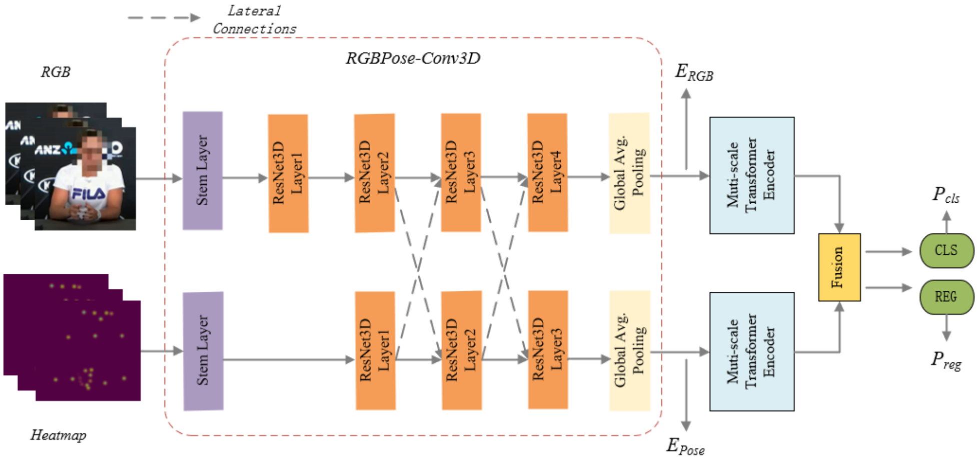

Figure 1: The framework of the proposed method, mainly consisting of RGBPose-Conv3D and Multi-scale Transformer Encoder.

hierarchical feature extractor, local Transformer and micro-gesture estimator. The overall architecture is shown in Fig. 1. Given a long sequence, the framework outputs the temporal positions (the starting and ending indexes in a long video sequence) and categories of micro-gestures. In the respective streams of RGB and heatmap, the motion features are firstly extracted from the long sequence by the RGBPose-Conv3D network, Then the extracted features are further processed into classification and regression branches by the multi-scale Transformer encoder with sharing weights, finally the spotting results of both modalities are fused simply by taking the average values to predict the interval and category of micro-gesture.

2.2. RGBPose-Conv3D

The effectiveness of micro-gesture online recognition hinges on the ability to detect subtle motion information in both spatial and temporal domains. Consequently, the choice of the backbone model is crucial in determining the detection performance. In the realm of image processing, it is a well-established practice that pre-trained classification models can be employed as the backbone for downstream tasks to extract features, such as object detection. Inspired by this, we opt for a video recognition model RGBPose-Conv3D [8] to serve as a backbone of our proposed methodology, which can process multi-modal data and represent the micro-gesture clip effectively. The structure of RGBPose-Conv3D as a feature extractor is shown in Fig.1.

Unlike the graph convolutional network, RGBPose-Conv3D is suitable for fusing multi-modal data such as RGB and Human skeletons. To feed 2D poses data into RGBPose-Conv3D, we reformulate them into a 3D heatmap volume at first. Then we pre-process the RGB and heatmap sequences and use a uniform sampling strategy [10] to ensure the consistency of the dual-stream inputs in the time dimension between the two training stages, and the RGB pathway has a smaller frame rate since RGB frames are low-level features. Then a two-stream 3D-CNN with two pathways is utilized to respectively process RGB modality and pose modality. Concretely, we first use a stem layer consisting of a 3D convolutional layer and a max-pooling layer, and then we choose ResNet50 as the backbone to extract 3D features from dual modalities. The two pathways are asymmetrical due to the different characteristics of the two modalities, the pose pathway has a smaller channel width, a smaller depth, as well as a smaller input spatial resolution compared to the RGB pathway. To avoid overfitting, RGBPose-Conv3D is trained on the SMG dataset in the first stage, which uses the sequence clips of the recognition task (Track 1) to learn the parameters, with two individual cross-entropy losses respectively for each pathway. The trained model is then utilized as a one-stage feature extractor to extract the motion-aware features, which can be matched with various micro-gesture spotting networks to achieve precise micro-gesture location. Given a pre-segmented clip with T frames, the feature extracted by the trained RGBPose-Conv3D with the input $I \in \mathbb{R}^{T \times C \times H \times W}$ can be embedded as $E \in \mathbb{R}^{T \times D}$, where H, W, and C represent the resolution and number of channels of RGB or heatmap frames, and D represent the features' dimensions of the two modalities, respectively. To address the problem of sequences with varying lengths, we opt for a fixed-length sliding window approach. By concatenating the features from various short segments along the temporal axis, we create a composite representation of motion features within each sliding window which are then fed into the subsequent processing modules. The concatenation process can be described as follows:

$$E_{sum} \ = \ Concatenation\left(E_1, E_2, \cdot\cdot, E_n\right), \tag{1}$$

where n is the number of clips in a sliding window, E_{sum} can be embedded as $E_{sum} \in \mathbb{R}^{(n \times T) \times D}$, and E_{RGB} and E_{Pose} in the Fig. 1 respectively represent the features E_{sum} extracted from two modalities.

2.3. Multi-scale Transformer Encoder

The extracted features of two modalities are further processed into classification and regression branches by two multi-scale Transformer encoders with sharing weights, which can be divided into three parts: hierarchical feature extractor, local Transformer and micro-gesture estimator. More complete details can be found in [9].

In the context of a sequence of micro-gestures with different durations, a hierarchical feature pyramid is advantageous for capturing temporal windows of varying lengths, thus obtaining multi-scale information. Notably, we have adopted 1D convolution with a kernel size of 3 x 1 and a stride of 1, followed by layer normalization and the SiLU activation function. To ensure the model's capability to detect micro-gestures with brief durations, the stride of the first layer in the feature pyramid is configured to 1, while subsequent layers are set to 2. Consequently, we derive multi-scale features F^i through processes of linear upsampling and concatenation. These operations facilitate the integration of rich contextual information and enhance the representational power of the features. Given a feature $E_{sum} \in \mathbb{R}^{T' \times D}$ extracted from the previous module, F^i can be embedded as $\{F^i \in \mathbb{R}^{(T'/2^{i-1}) \times D}, i = 1, 2, 3, 4\}$.

Given that micro-gestures are frequently inseparable with their surrounding frames, we have incorporated an attention mechanism to measure the similarity between frames and model the inter-dependencies among them. Instead of using a traditional Transformer [11] with a global attention mechanism, which may not be ideal for lengthy sequences, we have opted for a local Transformer by limiting attention within a local window [9], as it is more apt for our needs. This is because the temporal context beyond a specific range tends to be less informative for detecting micro-gestures, and global attention can lead to the redundancy of information that can hinder analysis. We generate a series of overlapping local windows along the time dimension of the multi-scale features F^i. Within each of these windows, we calculate self-attention which are then concatenated along the time dimension, to obtain a comprehensive representation of the micro-gesture sequence.

Given $F^i \in \mathbb{R}^{T \times D}$ from the hierarchical feature extractor, we generate encoded representations of Query(Q), Key(K), and Value(V) by projecting F^i through weight matrices $W_Q \in \mathbb{R}^{D \times D_Q}$, $W_K \in \mathbb{R}^{D \times D_K}$, $W_V \in \mathbb{R}^{D \times D_V}$, with the calculations as follows:

$$Q = F^i \times W_Q, K = F^i \times W_K, V = F^i \times W_V. \tag{2}$$

Then we apply Multi-head attention (MHA) within a local window through a defined series of operations:

$$\begin{array}{c} MHA(Q_i, K_i, V_i) = Concatenation(head_0, \ldots, head_n)W^O, \\ head_i = soft\max\left(\frac{Q_i K_i^T}{\sqrt{D_q}}\right) V_i \end{array}, \tag{3}$$

where n denotes the number of attention heads, W^O is the parameter matrix for the output of heads, Q_i, K_i and V_i respectively represent Q, K and V in the i-th local window. Finally the results of each window MHA are concatenated along the time dimension to generate the final encoded results, calculated by:

$$Y = \sum_i Concatenation\left(MHA\left(Q_i, K_i, V_i\right)\right), \tag{4}$$

where $Y \in \mathbb{R}^{T \times D}$ and $Concatenation\left(\cdot\right)$ denotes the operation of concatenating the MHA results along the time dimension.

The local Transformer's encoded features are subsequently processed by the estimator module to predict the precise location and category of micro-gestures. The estimator module is comprised of two distinct branches: a regression branch and a classification branch. The regression branch is tasked with estimating the temporal distance to the beginning and end frames of the micro-gesture at each point along the time dimension, while the classification branch is responsible for determining the specific category of the gesture.

To extract features related to both classification and regression, we utilize a channel attention mechanism, which is applied before sending the features to the head. To reduce overfitting, we implement weight sharing among these attention layers. For the purpose of accurately locating the interval of the gesture, which is crucial for micro-gesture detection, we adopt the strategy proposed by [12]. This strategy treats the regression challenge as a distribution prediction task, thereby accounting for the inherent uncertainty.

Given an encoded feature $Y \in \mathbb{R}^{T \times D}$, where T is the temporal dimension and D is the feature dimension, the output of the regression branch can be formulated as $P_{\text{reg}} \in \mathbb{R}^{T \times 2}$, and the output of the classification branch as $P_{\text{cls}} \in \mathbb{R}^{T \times C}$, with C is the total number of micro-gesture categories. In the subsequent phase of model training, both the local Transformer and the estimator module are trained conjointly using the training data of online recognition.

3. Experiments

3.1. Dataset and Metric

Dataset. The Spontaneous Micro-Gesture (SMG) dataset [1] is employed to evaluate our proposed method. The dataset consists of 3692 samples of 17 MGs, comprising of 40 sequences from 40 participants with an average age of 25 years. Each sequence lasts for around 15 minutes and the total dataset comprises 821,056 frames which exceeds 8 hours of recorded data. The participants are recorded and collected by Kinect resulting in four modalities, RGB, 3D skeletal joints, depth and silhouette. The dataset utilizes skeleton data from 25 human body joints, with each keypoint represented by 11 numerical values. In the MiGA2024 challenge, only RGB data which has a resolution of 1920×1080 and 2D skeletal points are allowed to be used as model input.

Metric. The definition of a true positive (TP) for each interval within a sequence is grounded in the overlap between the spotted interval and the ground-truth interval. The spotted interval $W_{spotted}$ is considered as TP if it satisfies the following condition:

$$\frac{W_{spotted} \bigcap W_{groundTruth}}{W_{spotted} \bigcup W_{groundTruth}} \geq k, \tag{5}$$

where k is set to 0.3, and $W_{groundTruth}$ denotes the ground truth of the micro-gesture interval (onset-offset). The performance of the model is then assessed using the F1-score, which is calculated as:

$$F1 - score = \frac{2TP}{2TP + FP + FN}, \tag{6}$$

where FP and FN represent the false positive and false negative, respectively.

3.2. Implementation Details

The RGBPose-Conv3D model is initially trained on micro-gesture data that has been pre-segmented from the SMG dataset, undergoing a total of 100 training epochs. Then the fully connected layer of the RGBPose-Conv3D model is removed, and the model is utilized as a feature extractor for our detection network. The duration of each video clip processed by the feature extractor is 8 frames, with an overlapping ratio 0.25 between clips. We configure the sliding window's length to be 512 frames.

In the local Transformer component of multi-scale Transformer encoder, the size of the local window is set to 8, and the overlap between windows is 4. Following this, the multi-scale Transformer encoder is secondly trained which lasts for 128 epochs. This training uses a cosine learning rate schedule and includes 5 warmup epochs. The Adam optimizer is employed with an initial learning rate set to 1×10^{-4}. The mini-batch size for training is 32, and the weight decay rate is 5×10^{-4}. To eliminate duplicate detection boxes and refine the results, we apply the Non-Maximum Suppression (NMS) [13] technique.

Table 1
The top-3 results of micro-gesture online recognition on the SMG dataset.

Rank	Team	F1-score
1	**Ours**	**0.27571**
2	HFUT-VUT	0.14346
3	JDY203	0.09289

Table 2
The results of various modality and different fusion strategies with different sliding window size and clip length.

clip length	sliding window size	Modality	Mode of fusion	F1-score
8	256	RGB	\	0.1835
8	512	RGB	\	0.1777
8	256	skeleton	\	0.2269
8	512	skeleton	\	0.2166
16	512	skeleton	\	0.2001
32	512	skeleton	\	0.1473
8	**256**	**RGB+skeleton**	**Late**	**0.2490**
8	**512**	**RGB+skeleton**	**Late**	**0.2757**
8	256	RGB+skeleton	Early	0.2644

3.3. Experimental Results

We firstly report the top-3 results of MIGA2024 Track2 to compare our best result with others' and show the effectiveness of our proposed method, as shown in Table 1. From the results, it can be observed that the designed network based on dual-stream input achieves promising performace among all methods.

We then validate the effectiveness of the proposed late fusion method of multimodal data by using data from only one modality RGB or skeleton for comparative experiments. Additionally, inspired by SlowFast [14], we add bidirectional lateral connections between the two pathways in RGBPose-Conv3D which is shown in Fig.1 to implement early-stage feature fusion between two modalities, and the early fusion features with the best performance on classification are sent into multi-scale Transformer for detection. In experiments, we find that early-stage feature fusion can also achieve improvement compared to using single modality.

We also explore the influence on the micro-gesture spotting of length of sliding windows and duration of video clips processed by the feature extractor. All the experiments results are shown in Table 2, it can be seen that small length of sliding windows and clips results in better performance under the condition of single modality, which maybe due to the larger number of clips for training and the reduction of redundant information within each clip. However, in the case of late fusion with multi-modal features, the smaller sliding window may be not conducive to the full fusion of information in two modalities.

4. Conclusion

In this paper, we proposed a deep network for dual-stream micro-gesture online recognition based on multi-scale Transformer. Our proposed method achieved excellent performance on the SMG dataset, but it is crucial to recognize that the development of micro-gesture online recognition is still in its early stages and there remains much room for improvement in terms of detection accuracy.

Acknowledgments

This work is partly supported by the Natural Science Foundation of Chongqing (No. CSTB2022NSCQ-MSX0977), the Key Research and Development Program of Shaanxi (No. 2023-ZDLGY-12), and the Practice and Innovation Funds for Graduate Students of Northwestern Polytechnical University (No. PF2024066).

References

[1] H. Chen, X. Liu, X. Li, H. Shi, G. Zhao, Analyze spontaneous gestures for emotional stress state recognition: A micro-gesture dataset and analysis with deep learning, IEEE Int. Conf. Automatic Face and Gesture Recognition (2019) 1–8.

[2] A. Vinciarelli, M. Pantic, H. Bourlard, Social signal processing: Survey of an emerging domain, Image Vis. Comput. 27 (2009) 1743–1759.

[3] B. de Gelder, J. V. den Stock, H. K. M. Meeren, C. B. A. Sinke, M. E. Kret, M. Tamietto, Standing up for the body. recent progress in uncovering the networks involved in the perception of bodies and bodily expressions, Neurosci. Biobehav. Rev. 34 (2010) 513–527.

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

[5] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, H. Lu, Skeleton-based action recognition with shift graph convolutional network, Proc. Computer Vision and Pattern Recognition (2020) 180–189.

[6] Z. Liu, H. Zhang, Z. Chen, Z. Wang, W. Ouyang, Disentangling and unifying graph convolutions for skeleton-based action recognition, Proc. Computer Vision and Pattern Recognition (2020) 140–149.

[7] X. Guo, W. Peng, H. Huang, Z. Xia, Micro-gesture online recognition with graph-convolution and multiscale transformers for long sequence, in: MiGA@IJCAI, 2023.

[8] H. Duan, Y. Zhao, K. Chen, D. Lin, B. Dai, Revisiting skeleton-based action recognition, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2959–2968.

[9] X. Guo, X. Zhang, L. Li, Z. Xia, Micro-expression spotting with multi-scale local transformer in long videos, Pattern Recognit. Lett. 168 (2023) 146–152.

[10] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, L. V. Gool, Temporal segment

networks: Towards good practices for deep action recognition, in: European Conference on Computer Vision, 2016.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Proc. Int. Conf. Neural Inf. Process. Syst. 30 (2017) 1–11.

[12] H. Zhang, Y. Wang, F. Dayoub, N. Sunderhauf, Varifocalnet: An iou-aware dense object detector, Proc. Computer Vision and Pattern Recognition (2021) 8510–8519.

[13] A. Neubeck, L. V. Gool, Efficient non-maximum suppression, Proc. Int. Conf. Pattern Recognit. 3 (2006) 850–855.

[14] C. Feichtenhofer, H. Fan, J. Malik, K. He, Slowfast networks for video recognition, 2019 IEEE/CVF International Conference on Computer Vision (ICCV) (2018) 6201–6210.