Micro-gesture Online Recognition with Graph-convolution and Multiscale Transformers for Long Sequence

XuPeng Guo[1], Wei Peng[2], Hexiang Huang[1] and Zhaoqiang Xia[1,3,*]

[1]School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710129, China
[2]Department of Psychiatry and Behavioral Sciences, Stanford University, California 94305, USA
[3]Innovation Center NPU Chongqing, Northwestern Polytechnical University, Chongqing 400000, China

Abstract

Micro-gesture is becoming a fundamental clue of emotion analysis and achieves more attention in this field. The studies are mainly focused on the task of micro-gesture classification which predicts the categories of micro-gesture while no works have been reported for spotting the micro-gestures. As a preliminary step for classification, the micro-gesture online recognition (spotting) that predicts the temporal location and category has achieved limited attention. In this context, we propose a novel deep network for micro-gesture online recognition, which incorporates the graph-convolution and multiscale transformer encoders. Specifically, we utilize a graph-convolution based Transformer module to extract motion features of 2D skeleton sequences, which are then processed by a feature pyramid module to obtain hierarchical multiscale features. We further employ a local Transformer module to model the similarity between micro-gesture frames, and decouple the classification and regression branches to achieve accurate location and category. These Transformers are trained in a two-stage strategy and combined to perform the spotting. Our proposed method is validated on the iMiGUE dataset and has achieved the **first ranking** in the task of online recognition (Track 2) of the MiGA2023 Challenge.

Keywords

Micro-gesture online recognition, Graph convolution, Multiscale Transformer

1. Introduction

In daily interactions, the human usually rely on body gestures to perceive emotions, which plays a crucial role in facilitating communication and understanding between individuals. With the increasing demand for intelligent systems, such as robots and other human-computer interaction systems, the ability to recognize and respond to users' emotions based on their body gestures has become a critical component [1]. Among the body gestures, micro-gesture (MiG) is an involuntary reaction triggered by people's inner emotions. To differentiate from more overt behavioral gestures [2], such as waving hand, micro-gestures are often more subtle and

MiGA@IJCAI23: International IJCAI Workshop on Micro-gesture Analysis for Hidden Emotion Understanding, August 21, 2023, Macao, China.
*Corresponding author.
Xpg_57@mail.nwpu.edu.cn (X. Guo); wepeng@stanford.edu (W. Peng); huanghexiang@mail.nwpu.edu.cn (H. Huang); zxia@nwpu.edu.cn (Z. Xia)
0000-0003-0630-3339 (Z. Xia)

conscious actions, such as biting finger, which are performed while attempting to conceal real feelings. 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 emotion indicators.

The micro-gesture analysis by computer vision techniques has attracted much attention in recent years. Micro-gesture analysis can mainly be divided into two classes [4]: 1) the classification of body gestures and 2) temporal body gesture localization and recognition (online recognition or spotting) in long sequences. The relevant researchers are committed to the former task, which conducts the classification of the pre-segmented clips, and most of the advanced technologies can achieve quite promising performance [5, 6]. The latter task is to detect the temporal frames with micro-gestures from a sequence and recognize it. Currently, there is a lack of automatic approaches of spotting micro-gestures, highlighting the importance of developing an automated micro-gesture detection model. This would enable more accurate and efficient analysis of micro-gestures, which are crucial for understanding and interpreting human emotions.

In this paper, to locate and recognize the micro-gestures from a long skeleton sequence, we propose a deep network for detecting micro-gestures by integrating the graph-convolution and multiscale Transformer encoders. We utilize a graph-convolution Transformer module based on hypergraphs and hyperedges to extract motion features of 2D skeleton sequences. Then the hierarchical multiscale features are obtained by a feature pyramid module. We further employ a multiscale Transformer module to model the similarity between micro-gesture frames. The classification and regression branches are finally decoupled to achieve accurate location and category. These Transformers are trained in a two-stage strategy and combined to perform the spotting. The main contributions of this paper can be summarized as:

- We design a deep network for MiG online recognition for the first time, which integrates the graph-convolution and multiscale Transformer encoders.
- We explore a graph-convolution Transformer as a feature extractor and a combination of feature pyramid and local Transformer to locate MiGs, which are trained separately in a two-stage way.
- We achieve the first ranking in the Track 2 of MiGA2023 challenge for online recognition.

2. Methodology

2.1. Overall Architecture

For spotting the micro-gestures, we propose a Transformer based network, which mainly consists of four important components: graph-convolution Transformer, 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 sequence) and categories of micro-gestures. In the graph-convolution Transformer, motion features are extracted from the long sequence by the graph convolution on hypergraphs and hyperedges. Then, the extracted features are further processed by the

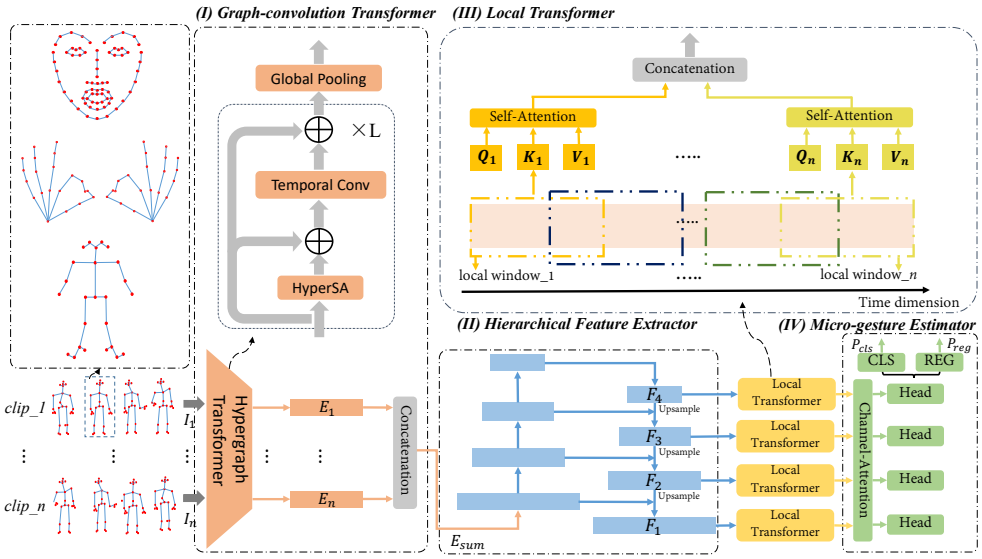

Figure 1: The framework of the proposed method, which mainly consists of four important components: (I) graph-convolution Transformer, (II) hierarchical feature extractor, (III) local Transformer, and (IV) micro-gesture estimator.

hierarchical feature extractor to obtain multiscale encoding features, which is fed to local Transformer for modelling the correlation between frames within an inner window. Finally, the interval of micro-gesture is predicted by decoupling classification and regression branches.

2.2. Graph-convolution Transformer

The performance of micro-gesture online recognition depends on the ability to capture subtle motion information in spatial and temporal dimensions. Therefore, the selection of backbone model plays a critical role in determining the detection performance. In the field of image processing, it is widely recognized that pretrained classification models can serve as the backbone of downstream tasks to extract features, such as object detection. Drawing inspiration from this, we also choose a video recognition model as the backbone for our proposed method. Although any graph-convolution network can be used as the backbone, in order to effectively process the 2D skeleton sequence, we choose to exploit the hypergraph Transformer [7] for the action recognition as the backbone to represent the micro-gesture clip, which is shown in Fig. 1 (I).

The use of hypergraphs and hyperedges in hypergraph Transformer (Hyperformer) allows for a more comprehensive representation of the input data, enabling the model to capture subtle nuances in the skeletal point relationships and structures that are crucial for accurate micro-gesture detection. Currently, hypergraph Transformer is primarily utilized for macro-action recognition, while there exists the significant difference between micro-gesture and behavioral action. To address this issue, we choose to train the Hyperformer on the iMiGUE dataset in the first stage, which uses the sequence clips of the recognition task (Track 1) to learn the parameters. 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 Hyperformer with the input $I \in \mathbb{R}^{T\times C\times N}$ can be embedded as $E \in \mathbb{R}^{(T/8)\times D}$, where N and D represent the number and characteristic dimensions of skeletal points, respectively. We choose to use fixed-length sliding window for solving the problem of varying sequence lengths. Therefore, we concatenate the features of different small fragments in the time dimension to obtain the motion features in a sliding window, which is fed into the subsequent module. The concatenation operation is given by:

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

where n is the number of clips in a sliding window, E_{sum} can be embedded as $E_{sum} \in \mathbb{R}^{(n\times T/8)\times D}$.

2.3. Hierarchical Feature Extractor

As the different durations of micro-gestures exist in a long sequence, hierarchical feature pyramid is beneficial to capture different temporal window lengths (multiscale information). The block module of the pyramid in our network shown in Fig. 1 (II) is similar to the C3 module in YOLOv5 [1] but with some key differences. Specifically, we utilize 1D convolution with kernel size of 3×1 and stride of 1, followed by layer normalization and SiLU activation function. In order to ensure that the model can capture micro-gesture features with a short duration, the stride of the first layer in the feature pyramid is set to 1, other layers are set to 2. Subsequently, we acquire multiscale features F^i through linear upsampling and concatenation operations, which enable the integration of rich contextual information and enhance the representation ability 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 $\{F^i \in \mathbb{R}^{(T^{'}/2^{i-1})\times D}, i = 1, 2, 3, 4\}$.

2.4. Local Transformer

Since the occurrence of micro-gesture is often inseparable from the contextual frames, we employ the attention mechanism to measure the similarity between frames and model the dependency of frames. Transformer [8] is utilized for similarity modeling between frames, but the traditional transformer with global attention mechanism may not be suitable for long sequence. It is recognized that the temporal context beyond a certain range is less informative for micro-gesture detection, and the global attention can introduce redundant information that interferes with the analysis. So we utilize the local Transformer by limiting attention within a local window [9], which is shown in Fig. 1 (III). A series of overlapping local windows are generated in the time dimension of F^i. Then we calculate self-attention in each window. Finally, the embedding results of each window are concatenated in the time dimension to obtain a comprehensive representation of the micro-gesture sequence.

Given $F^i \in \mathbb{R}^{T\times D}$, F^i is utilized to project encoded representations of Query(Q), Key(K), and Value(V) by using $P_Q \in \mathbb{R}^{D\times D_Q}$, $P_K \in \mathbb{R}^{D\times D_K}$, $P_V \in \mathbb{R}^{D\times D_V}$, which are given by:

$$Q = F^i \times P_Q, K = F^i \times P_K, V = F^i \times P_V \tag{2}$$

[1] https://github.com/ultralytics/yolov5

Then multi-head attention (MHA) will be applied in a local window by the following operations:

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

where n is the number of heads, W^O is the parameter matrix, Q_i, K_i and V_i respectively represent Q, K and V in the i-th local window. Then the results of each window MHA are concatenated in the time dimension to obtain the encoded results, which is given 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 $Concatenation\left(\cdot\right)$ the concatenation of MHA results in the time dimension.

2.5. Micro-gesture Estimator

Finally, the encoded features of local Transformer are fed to the estimator module to predict the location and category of micro-gestures. The estimator module consists of decoupling regression and classification branches, which is shown in Fig. 1 (IV). The former predicts the distance to the starting and ending frames of the micro-gesture at each point in the time dimension, while the classification branch is responsible for identifying the category to which it belongs. In order to obtain classification and regression related feature information, we employ the channel attention mechanism and apply it before sending the features to the head. To prevent overfitting, we enforce weight sharing among these attention layers. To further achieve accurate localization of the gesture interval for micro-gesture detection, we adopt the approach proposed by [10], which treats the regression problem as a distribution prediction problem to model uncertainty. Given an encoded feature $Y \in \mathbb{R}^{T\times D}$, the output of the regression branch can be embedded as $P_{reg} \in \mathbb{R}^{T\times 2}$, and the classification branch can be embedded as $P_{cls} \in \mathbb{R}^{T\times C}$, where C is the number of categories of micro-gestures. In the second stage of model learning, the local Transformer and estimator are jointly trained by the training data of online recognition.

3. Experiments

3.1. Dataset and Metric

Dataset. Micro-Gesture Understanding and Emotion analysis (iMiGUE) dataset [11] is employed to evaluate our proposed method. The dataset consists of 32 categories from post-match press conference videos of famous tennis players. The micro-gestures are annotated from 359 long video sequences and are captured in RGB modality and 2D skeletal joints collected from the Open-Pose algorithm. The 2D skeletal joints consist of a total of 137 key points, including 25 body points, 42 hand points, and 70 face points. In the MiGA2023 challenge, only 2D skeletal points are allowed to be used as model input.

Metric. The true positive (TP) per interval in one sequence is defined based on the intersection between the spotted interval and the ground-truth interval. The spotted interval $W_{spotted}$ is considered as TP if it fits the following condition:

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

where k takes 0.3, and $W_{groundTruth}$ represents the ground truth of the micro-gesture interval (onset-offset). F1-socre is then used to evaluate the performance of the model, which is given by:

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

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

3.2. Implementation Details

The hypergraph Transformer model is firstly trained using pre-segmented micro-gesture data from the iMiGUE dataset with a total of 200 epochs trained. Then, the fully connected layer of hypergraph Transformer is removed, and the model is utilized as the feature extractor for detection network. In Fig. 1 (I), the number of layers L in graph-convolution Transformer is 10. The length of one clip fed into feature extractor is 8, the overlap value is 2. We set the length of the sliding window to 512. In local Transformer, the local window size is set to 8, the overlap value is set to 4. The local Transformer and estimator are secondly trained for 200 epochs with a cosine learning rate schedule and 5 warmup epochs. We use Adam as an optimizer, where the initial learning rate is $1e-4$. The mini-batch size is 32, and the weight decay is 5e-4. None-Maximum Suppression (NMS) [12] is used to remove the duplicated boxes and obtain real results.

3.3. Experimental Results

As no baseline approach has been reported in the past, we only report the final performance on iMiGUE to evaluate the influence of two important components, i.e., the hypergraph Transformer and multiscale Transformer. Table 1 presents the performance of using other components to replace the above two components for micro-gesture online recognition. When HD-GCN [13] trained on iMiGUE is used as the feature extractor to observe the impact of hypergraph Transformer, the result of using HD-GCN indicates that this model achieves worse performance as it may be unable to accurately capture motion information. Subsequently, LSSNet [14], which has demonstrated strong performance in micro-expression detection, is selected as the detection network to observe the impact of multiscale Transformer. It also suggests that our model has more effectiveness in interframe modeling.

The visualization results of micro-gesture online recognition are demonstrated through two sets of skeletal sequences in Fig. 2. The first set shown in Fig. 2 (a) displays accurately spotted micro-gesture, while the second set depicted in Fig. 2 (b) demonstrates micro-gesture that is spotted incorrectly with an IoU value of 0.15. The results show that our method still faces challenges in accurately locating some samples.

Table 1
The performance of using various modules.

Hypergraph Transformer	Multiscale Transformer	Replacement	F1-score
✓	✗	LSSNet	0.0797
✗	✓	HD-GCN	0.0585
✓	✓	-	**0.1485**

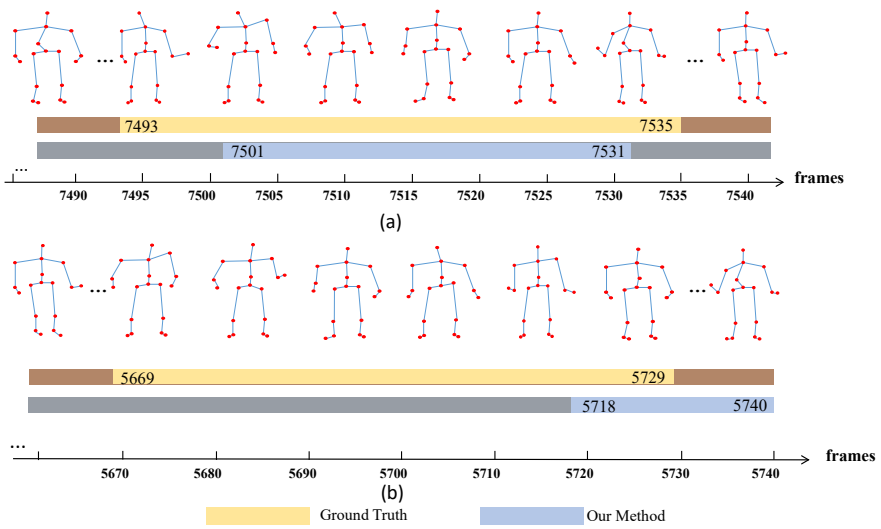

Figure 2: Visualization examples of micro-gesture online recognition by our proposed method.

4. Conclusion

In this paper, we proposed a Transformer based model for micro-gesture online recognition, which integrates graph-convolution and multiscale Transformers. Our proposed method achieved excellent performance on the iMiGUE dataset, but it is important to note that the development of micro-gesture online recognition is still in its early stages and there is 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), and the Key Research and Development Program of Shaanxi (Nos. 2021ZDLGY15-01 and 2023-ZDLGY-12).

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] Y. Zhou, C. Li, Z.-Q. Cheng, Y. Geng, X. Xie, M. Keuper, Hypergraph transformer for skeleton-based action recognition, arXiv abs/2211.09590 (2022).

[8] 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.

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

[11] 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, Proc. Computer Vision and Pattern Recognition (2021) 10626–10637.

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

[13] J. Lee, M. Lee, D. Lee, S. Lee, Hierarchically decomposed graph convolutional networks for skeleton-based action recognition, arXiv abs/2208.10741 (2022).

[14] W.-W. Yu, J. Jiang, Y.-J. Li, Lssnet: A two-stream convolutional neural network for spotting macro- and micro-expression in long videos, Proc. ACM Int. Conf. Multimedia (2021) 4745–4749.