

Micro-gesture Classification Based on Ensemble Hypergraph-convolution Transformer

Hexiang Huang^{1,†}, XuPeng Guo^{1,†}, Wei Peng² 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 classification has emerged as a significant research area within emotion analysis and human-computer interaction, garnering increasing attention. While some skeleton-based action recognition algorithms utilizing graph convolution networks have shown competence in micro-gesture classification, these deep models still face challenges in representing subtle temporal actions and handling the long-tailed distribution of samples. To address these issues, this paper proposes a deep framework with ensemble hypergraph-convolution Transformers, which fuses multiple models focused on various categories. In this model, the Transformers with hypergraph based attention are constructed and extended to enhance the representation ability of single model. Then a data grouping training and ensemble method is employed to handle imbalanced categories for micro-gestures, resulting in a significant improvement in classification accuracy of single models. Finally, our algorithm model is evaluated on the iMiGUE dataset, which achieves the Top-1 accuracy of **0.6302** and the **second ranking** in the MiGA2023 Challenge (Track 1: Micro-gesture Classification).

Keywords

Micro-gesture classification, Long-tailed distribution, Graph-convolution Transformer, Ensemble model

1. Introduction

Micro-gesture (MiG) classification refers to the process of identifying and categorizing small and subtle movements appeared on the human face and body, such as eye blinks, facial expressions, or hand gestures. The goal of automatic MiG classification is to accurately recognize and interpret these subtle movements, which can provide valuable insights into the understanding of a person's thoughts, emotions, and intentions. Deep learning algorithms and computer vision techniques [1, 2] are commonly employed to accomplish this task, finding frequent application in areas like human-computer interaction, emotion recognition, and biometric identification.

Due to the progress of deep learning techniques for action recognition [3, 4], the deep models have also been utilized to recognize the categories of MiG with the data of RGB and skeleton

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


*Corresponding author.

†These authors contributed equally.

✉ huanghexiang@mail.nwpu.edu.cn (H. Huang); Xpg_57@mail.nwpu.edu.cn (X. Guo); wepeng@stanford.edu (W. Peng); zxia@nwpu.edu.cn (Z. Xia)

ORCID 0000-0003-0630-3339 (Z. Xia)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

modalities. In the early works for MiG [5, 6], the RGB based methods, e.g., the temporal segmentation network (TSN) [7], and the skeleton based methods, e.g., spatio-temporal graph convolution network (ST-GCN) [8], originally for action recognition have been applied to evaluate the performance of recognizing MiGs as benchmarks. Although the RGB modality can provide more information of MiGs, the identity privacy for the people restricts the application of RGB modality. Therefore, the study focused on MiG tasks from skeleton modality.

Despite the dataset having been released for two years, there are currently limited reported works on skeleton-based MiG classification as the challenges of modeling subtle motions from the skeleton. But the graph convolutional networks (GCN) are commonly used in the task of skeleton based action recognition [9]. GCN is a graph-based presentation learning method originally designed for key point classification tasks. In applications, the relationships between different types of key points and edges in the graph need to be modeled, and these relationships can be very complicated. In this case, using a standard graph structure becomes less appropriate because high-order semantic correlation can be far more complicated than the binary relationships model by such graph. In contrast, hypergraphs [10, 11] provide more flexible and rich representation capabilities, which can be used to represent multiple relationships between key points of different types. The hyperedges can be used to construct complex relationships between key points of different types. By mapping the key points and edges in the hypergraph to a low-dimensional vector space, the graph neural network can not only improve its training capabilities but also enhance reasoning processes, thus provides a GCN with stronger and more comprehensive representation capabilities. So in order to capture the potential relationships that exist between the key points of human skeleton, the self-attention (SA) based on hypergraph [12] (called HyperSA) in a Transformer encoder was proposed to combine the Transformer [13] with skeleton for measuring both paired and higher-order relationships and applied to skeleton-based action recognition.

To capture the complicated relationships between different skeleton points from the face and body for MiGs, we extend the HyperSA by enhancing the self-attention weight with considering the relationship of hyperedges, which reorganizes the four parts of the SA module into different branches. These branches are integrated during the learning process, and the results obtained from this integration address the issue of insufficient learning from a single branch. Furthermore, since the data collected from real-world scenario often exhibit an imbalanced distribution, or a long-tailed distribution, a single model trained on relatively unbalanced data tends to exhibit biased predictions favoring the head categories, resulting in poorer performance on the tail categories. To overcome this problem, inspired by the data partitioning concept proposed by Cai et al. [14], we propose to partition the training data into two overlapping subsets and ensemble several independent models together by training them separately. The main contributions of this paper can be summarized as:

- We design a deep framework of ensemble hypergraph-convolution Transformer (EHCT) for the task of MiG classification.
- We extend the HyperSA by enhancing the hyperedges of SA module to promote the representation ability for MiGs.
- We leverage the ensemble strategies to combine several independent models to weaken the impact of imbalanced data.

- We perform extensive experiments and achieve the second ranking in the Track 1 of MiGA2023 Challenge.

2. Methodology

The main framework of our proposed method (EHCT) is shown in Fig. 1 (a). In the framework, we design two classifiers, namely, the main classifier and auxiliary classifier, by using the same-architecture base model of hyperformer convolution Transformer (HCT) to promote the discrimination ability and mitigate the long-tailed distribution of data. For the base model, the attention weight between key points and hyperedges are enhanced (eHyperSA) by considering the relationships between individual key points in the body, face, left and right hands. The details of three important components are described in the following section.

2.1. Hypergraph-convolution Transformer

As shown in Fig. 1 (b), the self-attention layer combined with the temporal convolution layer in the HCT is the basic block and stacked by L layers [12]. The skeletal input $S^T = \{\vec{s}_1, \vec{s}_2, \dots, \vec{s}_{137}\}$ comprises the key points extracted from a single frame, including those pertaining to the body, face, left and right hands, are presented in 2D format $\vec{s}_i = (x, y, c)$ by using the protocol of OpenPose [15]. According to self-attention mechanism [16], a linear transformation is applied to input S through multiplication with three weight matrices, resulting in the derivation of matrices Q , K , and V .

In the self-attention module of HCT shown in Fig. 1 (c), the feature E_f with the hyperedges of hypergraph is constructed by Eq. 1:

$$E_f = HD_e^{-1}H^T SW_e, \quad (1)$$

where H represents the incidence matrix of key points and hyperedges. In the matrix H , each row represents a key point and each column represents a hyperedge. D_e is the diagonal matrix representing the degree matrix of hyperedges, and W_e represents the projection matrix of hyperedges. Based on the hyperedge feature E_f , we extend the self-attention in our model (eHyperSA), which is expressed as follows:

$$A = \underbrace{QK^T}_a + \underbrace{QR_\phi^T}_b + \underbrace{QE_f^T}_c + \underbrace{E_f E_f^T}_d, \quad (2)$$

In the eHyperSA, the basic attention (components a and c) and relative positional embedding (component b) R_ϕ are used and similar to [17, 12]. In contrast to the vanilla HyperSA [12], the component d in the above equation is newly added, which considers the inner product of the hyperedge feature matrix E_f , improving the attention between hyperedges.

2.2. Main Classifier

Given that single base model may potentially impact the weight of components a and c in Eq. 2, in the interest of enhancing their weight and efficacy in the classification task, the main classifier explores multiple base models (HCTs) as multiple branches and integrates them directly.

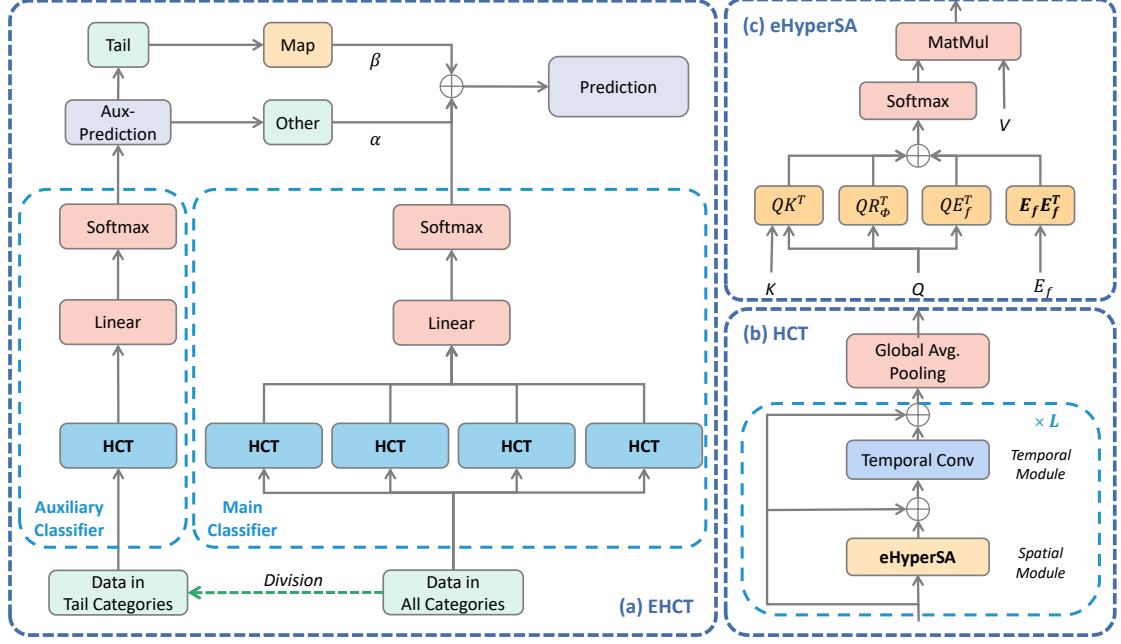


Figure 1: The overall framework of the proposed method: (a) ensemble hypergraph-convolution Transformer (EHCT), (b) hypergraph-convolution Transformer (HCT) module, (c) edge-enhanced hypergraph Self-Attention (eHyperSA) module.

To execute multi-branch integration, each branch in the main classifier emphasizes the primacy of components a and c while selectively incorporating components b and d . The corresponding mathematical equation for attention in each branch is show in Eq. 3:

$$A_i = QK^T + QE_f^T + \sum_i C_2^i \{QR_\phi^T, E_f E_f^T\}. \quad (3)$$

The matrix V is multiplied by the output of each branch's attention calculated by the Softmax function, and the integration is performed as follows:

$$B_i = ReLU(TemporalConv(Softmax(A_i)V_i)), \quad (4)$$

The final output is obtained by taking the average of the output logits B_i of each branch, which is shown in Eq. 5:

$$Logits_{main} = \frac{\sum_i^{\mathcal{N}} B_i}{\mathcal{N}}, \quad (5)$$

where the parameter \mathcal{N} denotes the number of branches.

2.3. Auxiliary Classifier

The data used in the task of MiG classification usually exhibit an imbalanced distribution across the different categories with a long-tailed distribution, e.g., the iMiGUE dataset [6]. In order

to mitigate the adverse impact of the sample imbalance, inspired by GoogLeNet [18] and ACE [14], we design an auxiliary classifier.

The data are bifurcated based on the count of data instances per category into two major categories, namely the head and tail categories. Subsequently, all data instances corresponding to the tail categories are extracted, and the same number of instances as the tail categories are randomly selected from the head categories to form the tail training set. In this tail training set, the labels of the selected instances from the head categories are reassigned to other categories, while the labels of the tail categories are one-to-one mapped to the original labels in the dataset.

With the logits from the main classifier and the auxiliary classifier, the way of combining these two outputs is calculated as follows:

$$Logits = Logits_{main} + \alpha \cdot Other \{Logits_{aux}\} + \beta \cdot Map \{Logits_{aux}\}, \quad (6)$$

where the hyperparameter α denoted as the weight by which the logits of the auxiliary classifier, when predicted as the other category, is accumulated into the logits of the main classifier, and the hyperparameter β denoted as the weight by which the logits of the auxiliary classifier, when predicted as a tail category, is accumulated into the logits of the main classifier through a mapping relationship. The final prediction can be obtained from the following equation:

$$Prediction = argmax_i (Logits(i)). \quad (7)$$

3. Experiments

In this section, we evaluate our model on the iMiGUE dataset [6] by following the protocol of MiGA2023 Challenge (Track 1: Micro-gesture Classification). The dataset, metrics, ablation study and comparison experiments are reported in the following sections.

3.1. Dataset and Metrics

In this challenge, the iMiGUE [6] dataset with fixed training and test samples is used to evaluate our proposed method. This dataset includes a total of 32 categories of MiGs, and covers two emotions as well as 72 subjects with each gender accounting for half of the total number of subjects. It consists of 18,499 samples taken from 359 videos with a resolution of 1280×720 . Each video is about 0.5-25.8 minutes long. Since the iMiGUE dataset is collected in-the-wild setting, the overall dataset presents a long-tailed (unbalanced) distribution.

To evaluate the classification performance of our model, we employ Top-1 accuracy and Top-5 accuracy as evaluation metrics, the equations of the metrics are as follows:

$$Acc_{Top-1} = \frac{\sum_{i=1}^N [argmax(P(y_i|x_i)) = y_i]}{N}, \quad (8)$$

$$Acc_{Top-5} = \frac{\sum_{i=1}^N [y_i \in top5(P(y_i|x_i))]}{N}, \quad (9)$$

where N denotes the number of samples, x_i denotes the feature of the i -th sample, y_i denotes the true label of the i -th sample, $P(y_i|x_i)$ denotes the probability distribution obtained from

Table 1

Performance comparison using various components on iMiGUE dataset, where a , b , c , d denote the components in Eq. 2, d_{ori} represents the original component d in vanilla HyperSA, B1~B5 denote single base models, and E1~E3 denote ensemble models.

Method	Self-Attention				Ensemble	Aux-classifier	Accuracy(%)	
	a	b	c	d			Top-1	Top-5
Hyperformer [12]	✓	✓	✓	d_{ori}	✗	✗	57.01	87.86
B1: Ours ($a + c + d_{ori}$)	✓	✗	✓	d_{ori}	✗	✗	58.35	87.88
B2: Ours ($a + c$)	✓	✗	✓	✗	✗	✗	57.83	89.22
B3: Ours ($a + b + c$)	✓	✓	✓	✗	✗	✗	58.57	89.43
B4: Ours ($a + b + c + d$)	✓	✓	✓	✓	✗	✗	58.79	89.11
B5: Ours ($a + c + d$)	✓	✗	✓	✓	✗	✗	58.09	90.27
E1: Ours (B3+B4+B5)	✓	✓	✓	✓	✓	✗	60.68	89.00
E2: Ours (B2+B3+B4+B5)	✓	✓	✓	✓	✓	✗	61.38	90.22
E3: Ours (B2+B3+B4+B5)	✓	✓	✓	✓	✓	✓	63.02	91.36

the model’s predictions for the i -th sample, and $top5$ denotes the top five categories with the highest probabilities.

In our experiments, the key parameter settings are configured as follows: 150 training epochs, a batch size of 8, an initial learning rate of 0.0005, and a learning rate decay rate of 0.1. In Fig. 1 (b) HCT, the number of stacked layers L is set to 10. All experiments are performed on an NVIDIA GeForce RTX 4090.

3.2. Ablation Study

Firstly, in order to verify the effectiveness of various parts of the self-attention mechanism based on the skeletal structure of human body, we conduct a series of ablative research experiments, and the specific results can be obtained from Table 1.

We use the vanilla Hyperformer model as the baseline and remove the relative position encoding b and bias d_{ori} for the four components of the attention module in vanilla HyperSA to observe the role of each component. Through the results, it can be observed that compared to the baseline, when we remove both the relative position encoding b and bias d_{ori} , the Top-1 accuracy is improved by 0.82%. When we remove only the relative position encoding b or bias d_{ori} separately, the Top-1 accuracy is improved by 1.34% and 1.56%, respectively. Therefore, we believe that the relative position encoding b and bias d_{ori} in HyperSA may not have verify significant effects on attention extraction.

Next, in order to further improve the accuracy of the model, we improve the original bias d_{ori} into the current component d , which is the attention between hyperedges obtained through the inner product of hyperedge features. By doing this, the Top-1 accuracy of the model is increased by 1.78% compared to the baseline, indicating that the attention between hyperedges has achieved significant effects on MiGs.

Due to the phenomenon of overfitting that may occur during the training of one single model, its performance may be good only on the training set, but it may decrease when facing new data. Moreover, when the dataset is complex, a single model often cannot learn global patterns.

Table 2

The comparison results of various methods on iMiGUE dataset.

Methods	Model+Modality	Accuracy(%)	
		Top-1	Top-5
ST-GCN [8]	GCN + Skeleton	46.97	84.09
MS-G3D [19]		54.91	89.98
TSN [7]	2DCNN + RGB	51.54	85.42
TRN [20]		55.24	89.17
TSM [3]		61.10	91.24
Hyperformer [12]	Transformer + Skeleton	57.01	87.86
EHCT (Ours)		63.02	91.36

Therefore, we integrate multiple models using different attention components in training to improve the generalization and robustness of the single model.

We employ ensemble learning with three branches, which improves the Top-1 accuracy by 3.67% compared to the baseline. To further enhance the attention weights between key points (component a) and between key points and hyperedges (component c), we add branches that only utilize components a and c , respectively, resulting in a four-branch ensemble approach. This further improves the Top-1 accuracy by 4.37% compared to the baseline.

Furthermore, we select all categories with instance counts at 1/50 of the maximum instance count as the tail categories. At the same time, we select head categories with a ratio of approximately 1:1 to merge with the tail categories and construct an independent training set. Through this training set, an auxiliary classifier is trained that uses all components of attention. The model with the auxiliary classifier achieves a Top-1 accuracy of 63.02%, which is an improvement of 6.01% compared to the baseline.

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

Our proposed technique is also examined through a comparative analysis on iMiGUE dataset, which is shown in Table 2. We compare our proposed method with state-of-the-art methods such as 2D convolutional networks utilizing CNNs with RGB data, GCNs with skeleton data, and Transformers with skeleton data. Compared to the MS-G3D [19] method, which utilizes 3D GCN on skeleton data, our method demonstrates an improvement of 8.11% on Top-1 accuracy and 1.38% on Top-5 accuracy. In comparison with the TSM [3] method, which employs 2D convolutional networks on RGB data, our method improves Top-1 accuracy by 1.92% and Top-5 accuracy by 0.12%. Compared to the Hyperformer [12] method, which uses Transformer on skeleton data, our method shows the significant improvement of 6.01% in Top-1 accuracy and 3.5% in Top-5 accuracy. It is observed that our proposed method (EHCT) outperforms the other methods, achieving the best performance on the iMiGUE dataset.

4. Conclusions

In conclusion, this paper introduces a deep framework that utilizes ensemble models based on hypergraph-convolution Transformer for the MiG classification from human skeleton data. The skeleton is organized by the proposed hypergraphs, which enable to capture complex correlations. By enhancing the attention mechanism and multimodel fusion techniques, the proposed method effectively extracts subtle dynamic features from different gestures. As a result, our designed model surpasses the state-of-the-art performance on the iMiGUE dataset, demonstrating its effectiveness in accurate classification of human skeleton data.

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] Z. Xia, W. Peng, H.-Q. Khor, X. Feng, G. Zhao, Revealing the invisible with model and data shrinking for composite-database micro-expression recognition, *IEEE Transactions on Image Processing* (2020) 8590–8605.
- [2] X. Guo, X. Zhang, L. Li, Z. Xia, Micro-expression spotting with multi-scale local transformer in long videos, *Pattern Recognit. Lett.* (2023) 146–152.
- [3] J. Lin, C. Gan, S. Han, Tsm: Temporal shift module for efficient video understanding, *International Conference on Computer Vision (ICCV)* (2019) 7082–7092.
- [4] W. Peng, J. Shi, Z. Xia, G. Zhao, Mix dimension in poincaré geometry for 3d skeleton-based action recognition, *ACM International Conference on Multimedia (ACM MM)* (2020) 1432–1440.
- [5] 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* (2023) 1346–1366.
- [6] 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, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2021) 10626–10637.
- [7] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, L. V. Gool, Temporal segment networks for action recognition in videos, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019) 2740–2755.
- [8] S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, *AAAI Conference on Artificial Intelligence (AAAI)* (2018) 7444–7452.
- [9] T. Ahmad, L. Jin, X. Zhang, S. Lai, G. Tang, L. Lin, Graph convolutional neural network for human action recognition: A comprehensive survey, *IEEE Transactions on Artificial Intelligence* (2021) 128–145.

- [10] Y. Feng, H. You, Z. Zhang, R. Ji, Y. Gao, Hypergraph neural networks, AAAI Conference on Artificial Intelligence (AAAI) (2019) 3558–3565.
- [11] S. Bai, F. Zhang, P. H. S. Torr, Hypergraph convolution and hypergraph attention, Pattern Recognition (2021) 107637.
- [12] 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).
- [13] C. Ying, T. Cai, S. Luo, S. Zheng, G. Ke, D. He, Y. Shen, T.-Y. Liu, Do transformers really perform bad for graph representation?, Neural Information Processing Systems (NeurIPS) (2021).
- [14] J. Cai, Y. Wang, J.-N. Hwang, Ace: Ally complementary experts for solving long-tailed recognition in one-shot, International Conference on Computer Vision (ICCV) (2021) 112–121.
- [15] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, Y. Sheikh, Openpose: Realtime multi-person 2d pose estimation using part affinity fields, IEEE Transactions on Pattern Analysis and Machine Intelligence (2018) 172–186.
- [16] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, Neural Information Processing Systems (NeurIPS) (2017).
- [17] K. Wu, H. Peng, M. Chen, J. Fu, H. Chao, Rethinking and improving relative position encoding for vision transformer, International Conference on Computer Vision (ICCV) (2021) 10013–10021.
- [18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015) 1–9.
- [19] Z. Liu, H. Zhang, Z. Chen, Z. Wang, W. Ouyang, Disentangling and unifying graph convolutions for skeleton-based action recognition, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 140–149.
- [20] B. Zhou, A. Andonian, A. Torralba, Temporal relational reasoning in videos, European Conference on Computer Vision (ECCV) (2018) 831–846.