The 2nd Challenge on Micro-gesture Analysis for Hidden Emotion Understanding (MiGA) 2024: Dataset and Results

Haoyu Chen[1], Björn W. Schuller[2], Ehsan Adeli[3] and Guoying Zhao[1,*]

[1]*CMVS, University of Oulu, Finland*
[2]*GLAM, Imperial College London, United Kingdom*
[3]*Stanford University, USA*

Abstract

This paper summarizes the 2nd Challenge of Micro-gesture Analysis for Hidden Emotion Understanding (MiGA) 2024. The competition was split into two independent tracks: micro-gesture classification from pre-segmented data clips, and micro-gesture online recognition in sequences of continuous data. In this edition of the MiGA challenge, both tracks use multi-modal data (RGB and skeleton as modalities). For evaluation, accuracy for classification and F1 score for online recognition are used as the evaluation measure. Two large micro-gesture datasets (iMiGUE and SMG) were made publicly available and the Kaggle platform was used to manage the competition. Results achieved a classification accuracy of 70.25% for micro-gesture classification, showing a significant improvement compared to last year's competition, meanwhile, an F1 score for online recognition is about 0.2757 was achieved for multi-modal gesture recognition, showing the task is still challenging and leaves considerable margin for improvement.

Keywords

Affective computing, behavior analysis, multi-modal gesture recognition, micro-gestures, emotion understanding

1. Introduction

Understanding emotions is fundamental to human intelligence and should hold a similar significance in artificial intelligence [1]. In the domains of emotion analysis and recognition, prior research has largely concentrated on facial expressions, vocal intonations, and physiological indicators such as heart rate. Relatively few studies, however, have explored the interpretation of emotions through gestural behaviors. Psychological research indicates that body language plays a crucial role in understanding emotions [2]. Notably, when individuals attempt to conceal their feelings, they often adjust their facial expressions but struggle to completely suppress micro-expressions. Moreover, only a few people actively manage their body movements in these

MiGA@IJCAI24: International IJCAI Workshop on Micro-gesture Analysis for Hidden Emotion Understanding, August 4th, 2024, Jeju, Korea.

*Corresponding author.

chen.haoyu@oulu.fi (H. Chen); bjoern.schuller@imperial.ac.uk (B. W. Schuller); eadeli@stanford.edu (E. Adeli); guoying.zhao@oulu.fi (G. Zhao)

0000-0003-3267-2664 (H. Chen); 0000-0002-6478-8699 (B. W. Schuller); 0000-0002-0579-7763 (E. Adeli); 0000-0003-3694-206X (G. Zhao)

situations. This suggests that gestures may offer valuable insights into hidden or suppressed emotions [3, 4, 5].

With the above observations, we initiated the workshop series focusing on the **micro-gesture (MG) analysis for hidden emotion understanding**, which is a novel research direction for the computer vision community. Micro-gestures (MGs) are subtle, involuntary body movements that can reveal suppressed or hidden emotions, often used in psychology to interpret inner feelings. MGs encompass various gestures – such as scratching the head, touching the nose, or fidgeting with clothing – and differ from typical gestures by lacking any communicative or illustrative purpose. Instead, they are spontaneous responses to certain stimuli, especially negative ones. Existing researchers mainly work on ordinary gestures/actions [6, 7], which are often used to convey semantic meanings or express attitudes. Meanwhile, MGs may occur when individuals try to mask true emotions, like stress or nervousness, in high-stakes situations. These micro-expressions can provide insight into hidden emotional states and may also correlate with neurological or mental disorders, making them valuable for diagnostic support. Automatic MG recognition has promising applications in areas such as human-computer interaction, social media, public safety, and healthcare.

In 2023, MiGA organized the first challenge on MG recognition with only skeleton modality data recorded with Kinect V2[1]. In the 2023 challenge, 54 entrants participated in the MiGA challenge which was devoted to skeleton-based MG classification and online recognition. With more research interests gained in the recognition of micro-gestures [8, 9, 10, 11], we chose to continuously host the MiGA competition this year. In the edition of 2024 this year [2], we have organized a second round of the same two tasks (classification and online recognition) including more modalities (both RGB and skeleton). Two public MG datasets (iMiGUE and SMG) [12, 13, 3] are used.

In this paper, we detail how the MiGA-IJCAI 2014 challenge was organized, the datasets, the results achieved by 72 entrants who joined the competition, and the implementing schemes of the winning methods.

2. Competition Tracks and Datasets

In this section, we introduce the two challenge tracks and their corresponding characteristic, as well as the datasets used in each track.

2.1. Track 1: Multi-modal MG classification

In this track, we focus on classifying micro-gestures (MGs) based on both skeleton data and RGB data from pre-segmented short video clips. This track utilizes an in-the-wild MG dataset, the iMiGUE dataset, which includes video footage of tennis players during post-match interviews, featuring detailed ground-truth annotations for MGs. Unlike typical action or gesture datasets, MGs capture finer, more subtle body movements that occur naturally in real-world interactions. Key challenges in this classification task include learning these intricate movement patterns,

[1] https://cv-ac.github.io/MiGA2023/

[2] https://cv-ac.github.io/MiGA2/

managing the imbalanced distribution of MG samples, and distinguishing the high variability in MGs across classes.

For this year's challenge, we also provide RGB data alongside the skeleton data and encourage participants to explore multimodal approaches that integrate both modalities. The training and testing sets are drawn from the iMiGUE dataset, following a cross-subject evaluation protocol: 72 subjects are split, with 37 subjects allocated for training and 35 for testing. Specifically, for MG classification, 13,936 clips are designated for training, 3,692 clips for validation, and an additional 4,563 clips are used for testing without annotations.

2.2. Track 2: Multi-modal MG online recognition

In this track, we tackle online micro-gesture (MG) recognition by utilizing both skeleton data and RGB data from extended video sequences. Unlike traditional online action or gesture recognition datasets, where actions are typically well-structured and sequentially organized, MG samples appear spontaneously and in varied sequences, resembling natural communicative behaviors. Consequently, the task of online MG recognition requires handling complex transitions between body movements, including the simultaneous occurrence of multiple MGs, partial or incomplete MGs, and intricate transitions. Additionally, detecting subtle MGs amidst other, less relevant body movements adds another layer of complexity, presenting challenges not commonly addressed in previous gesture recognition research.

The same as track 1, we encourage participants to adopt multimodal approaches, while the SMG dataset serves as the foundation for this track. A cross-subject evaluation protocol is applied, with sequences from 35 subjects allocated for training and sequences from the remaining 5 subjects reserved for testing.

3. Competition Itinerary

This section encompasses both the competition schedule and relevant participant details.

3.1. Competition agenda

The challenge was managed using the Kaggle competition framework. The schedule of the competition was as follows:
March 29, 2024. Call for Challenge online. Registration starts.
April 9, 2024. Release of training data, development toolkit, and sample codes.
May 2, 2024. Release of testing data.
May 12, 2024. Final testing data and result submission. Registration ends.
May 17, 2024. Release of challenge results.
May 30, 2024. Paper submission deadline (workshop).
June 04, 2024. June 07, 2024. Notification to authors.
June 04, 2024. June 12, 2024. Camera-ready deadline.
August 03 – 09, 2024. MiGA IJCAI 2024 Workshop, Jeju, Korea.

3.2. Participants

As stated, the competition has been conducted using Kaggle, a well-known challenge open-source platform. We created a different competition for each track, having separate information and leaderboard [3] [4]. A total of 72 users have been registered in the Kaggle platform, 43 for track 1 and 29 for track 2 (note that some users might have been registered for more than one track but we count each track severately). All these users were able to access the data for the developing stage and submit their predictions for this stage. For the final evaluation stage, team registration was mandatory, and a total of 16 teams were successfully registered: 12 for track 1, and 4 for track 2. During the challenge period, in total 323 submissions were made with 235 for track 1 and 88 for track 2.

4. Protocol and Evaluation

In this section, we introduce the evaluation metrics used to evaluate the participants for the two tracks.

4.1. Multi-modal MG classification

We evaluate participants' methods based on Top-1 accuracy in the challenge, but we encourage participants to report their results when submitting papers to our MiGA workshop on the following subsets of the test set: 1) Overall: All segments in the test split; 2) Tail Classes: Due to the long-tailed nature of the datasets, among a total of 33 classes in the iMiGUE dataset, 28 classes are tail classes (approx. 57.8 % of the data).

As to the submission format for classification track, participants must submit their predictions in a single .csv file. Instructions and sample submission files is released with the data. For each 'Id' in the validation set, they must predict a probability for the 'Target' variable. The file should contain a header (Id, Target).

4.2. Multi-modal MG online recognition

As for the MG online recognition track, we jointly evaluate the detection and classification performances of algorithms by using the F1 score measurement defined below: F1 =2Precision*Recall/(Precision+Recall), given a long video sequence that needs to be evaluated, Precision is the fraction of correctly classified MGs among all gestures retrieved in the sequence by algorithms, while Recall (or sensitivity) is the fraction of MGs that have been correctly retrieved over the total amount of annotated MGs.

Considering the submission format for online recognition track, participants must submit their predictions in a single .csv file. The submission .csv file should consist of the following columns with headers: ID: incremental index, class: prediction label, start_frame: staring frame, end_frame: ending frame, sample_id: represents the subject, i.e., the sample folder name is Sample0005, the sample_id is 5.

[3] https://www.kaggle.com/competitions/2nd-miga-ijcai-challenge-track1/

[4] https://www.kaggle.com/competitions/2nd-miga-ijcai-challenge-track2/

Table 1
Track 1 Multi-modal MG classification results in the iMiGUE dataset.

Team	Accuracy	Rank	Modality	Backbone
HFUT-VUT	70.25%	1	RGB+skeleton heat map	PoseConv3D
NPU-MUCIS	70.19%	2	RGB+skeleton + skeleton heat map	PoseConv3D Res2Net3D GCN
ywww11	68.92%	3	RGB+skeleton heat map	PoseConv3D CLIP

For both of the two tracks, the results are evaluated on the server and displayed on the ranking list in real time. The organization team has the right to examine the participants' source code to ensure the reproducibility of the algorithms. The final results and ranking are confirmed and announced by the organizers after verifying the reproducibility of the source code.

5. Challenge Results and Methods

In this section, we report the winning methods proposed by the participants. For the two tracks, we asked the top three teams to submit their source code and predictions for the test sets. Below, we introduce the implementation details of each method.

5.1. Track 1: Multi-modal MG classification

Table 1 summarizes the methods of the three teams who ranked the top three on the test set of track 1. One can see that all the methods use both RGB and skeleton modalities and tend to convert skeleton modality data into heat map presentations with a PoseConv3D [14] backbone. Next, we describe the main characteristics of the three winning methods.

First place: The HFUT-VUT team proposes the first place scheme [15] with the core structure of the proposed approach as two separate branches for RGB and Pose data. Initially, it employs the PoseConv3D network [14] as its backbone, optimizing the learning of spatio-temporal features and improving resilience against noise. Specifically, the method is built on a two-stream 3D CNN backbone, with the top pathway dedicated to RGB data processing and the bottom pathway handling skeleton data. A cross-attention fusion module is then introduced to capture interactions between the RGB and Pose modalities. Finally, drawing inspiration from [16], a prototypical refinement module is added. This module establishes prototype representations for each fine-grained micro-gesture class during training, prompting the model to refine ambiguous samples across different gesture categories. By jointly leveraging the PoseConv3D backbone for both the RGB and skeleton modalities, the model achieves 67.91% accuracy on the skeleton-only modality and 70.25% accuracy when combining RGB and skeleton data.

Second place: The NPU-MUCIS team introduces a framework named M2HEN [17], which constructs a heterogeneous ensemble network by combining two fundamentally different deep learning models: a 3D convolution-based model and a Transformer-based model. This

Table 2
Track 2 Multi-modal MG online recognition results on the SMG dataset.

Team	F1 score	Rank	Modality	Backbone
NPU-MUCIS	0.2757	1	RGB+skeleton heat map	PoseConv3D
HFUT-VUT	0.1435	2	RGB	I3D+Mamba

heterogeneous ensemble approach enhances feature diversity and strengthens the model's representational power. For the 3D convolutional model, they present the MiG-enhanced Multi-modal and Multi-scale 3D Convolutional sub-Network (M3CN), while the Ensemble Hypergraph-Convolution Transformer (EHCT) [18] is employed as the Transformer model. With either RGB or skeleton data alone, the framework achieves 61.14% and 61.11% accuracy, respectively. When fusing both modalities, accuracy increases to 66.57% (baseline) and further improves to 70.19% with the ensemble models and group training.

Third place: The ywww11 team presents a method rooted in the CLIP framework. Building on Froster CLIP [19], they introduce a token attenuation strategy within the video encoding module, which incrementally filters out less significant tokens at each layer. For the skeleton modality, they align it with the video modality's CLIP model by applying text embeddings from the video modality to the skeleton network. The PoseConv-3D model is specifically enhanced with CLIP text embeddings [14], allowing it to work in conjunction with the CLIP text encoder. This integration fosters collaborative processing. By fully leveraging feature extraction from both the skeleton and video modalities, this approach boosts performance in micro-gesture recognition tasks by using both skeleton sequences and video frames. By combining three modalities (RGB, Skeleton Joint, Skeleton Limb) with optimized weights, they reach an accuracy of 68.90%.

5.2. Track 2: Multi-modal MG online recognition

Table 2 summarizes the methods of the top two teams ranked on the test set of track 2. Since the third runner used exactly the same methodology as the baseline [20, 21], we do not report their method here. Next, we describe the main characteristics of the two winning methods.

First place: To detect micro-gestures using RGB and skeleton data, team NPU-MUCIS [22] proposes a network that is primarily composed of two key elements: a 3D convolutional network (RGBPose-Conv3D [14]) and a multi-scale Transformer encoder [23]. The Transformer encoder itself consists of three main parts: a hierarchical feature extractor, a local Transformer, and a micro-gesture estimator. For each long video sequence, the framework provides both the temporal positions (start and end indices) and the categories of the micro-gestures. Within the RGB and heatmap streams, motion features are initially extracted from the sequence using the RGBPose-Conv3D network. These features are then fed into classification and regression branches in the multi-scale Transformer encoder, which shares weights across branches. Finally, the detection results from both modalities are averaged to predict the intervals and categories of the micro-gestures. This approach achieves F1 scores of 0.1835 for RGB-only, 0.2269 for skeleton-only, and a combined score of 0.2757 when both RGB and skeleton modalities are used together.

Second place: The approach introduced by the team HFUT-VUT [24] is composed of a video encoder and an action decoder. For each video sequence, the model learns a series of trainable query points that help identify action boundary positions, along with query vectors that interpret action semantics and locations from the input features. The action decoder, incorporating a Mamba-MHSA module and a multi-level interaction module, then maps features to linear projection layers, which decode action labels from the query vectors and convert the query points into detection outputs. This method achieves F1 scores of 0.1835 for RGB-only, 0.2269 for skeleton-only, and 0.2757 when combining both RGB and skeleton data. With a single RGB modality, it achieves an F1 score of 0.1434.

6. Discussion

This paper has described the main characteristics of the MiGA 2024 Challenge hosted at IJCAI 2024 which included tracks on (i) Multi-modal MG classification, and (ii) Multi-modal MG online recognition. Two large datasets (the SMG and iMiGUE datasets) were introduced and made publicly available with corresponding toolkits to the participants for a fair comparison of the performance results.

Analyzing the methods introduced by the above participants, several conclusions can be drawn. For multi-modal MG classification (track 1), although a large improvement has been made in the performances from this year's method compared to last year's models [25, 18], there is still considerable room for improvement (the current highest accuracy is 70.25%). On the other hand, there are still many ways to improve in the MG online recognition task from multi-modal data. For instance, the detection performances are still quite low with the highest F1 score as 0.2757, showing that spotting MG is a challenging task to perform even by humans. Aside from those winning schemes proposed for the MiGA competition 2024, some other interesting research related to MG is also included in the MiGA 2024 workshop. For instance, Xia et al. propose to use event data to recognize micro-gestures and micro-expressions which is a novel research entry that meets the nature of those short and rapid movements of human behaviors [26].

Future trends in MiGA may include hidden emotion understanding via MGs with the analysis of social signals, and face expression analysis as relevant information cues. Besides, extending the datasets to larger scales and diverse materials toward more real-world scenarios can also be a promising direction.

Acknowledgments

We wish to acknowledge the teams who participated in MiGA 2024. We especially thank Associate Professor Xiaobai Li for assisting the organizing the event, and Marko Savic, Atif Shah, and Abdelrahman Mostafa for the technical support. We also wish to thank the platforms Codalab and Kaggle for providing free sources for us to organize the challenges of MiGA 2023 and 2024.

This MiGA challenge and workshop was supported by the Academy of Finland for Academy Professor project EmotionAI (grants 336116, 345122), the University of Oulu & The Academy of

Finland Profi 7 Hybrid Intelligence (grant 352788), by the Ministry of Education and Culture of Finland for AI forum project. We also wish to acknowledge the CSC – IT Center for Science, Finland, for computational resources.

References

[1] G. Zhao, Y. Li, Q. Xu, From emotion ai to cognitive ai, International Journal of Network Dynamics and Intelligence (2022) 65–72.

[2] J. Z. Wang, S. Zhao, C. Wu, R. B. Adams, M. G. Newman, T. Shafir, R. Tsachor, Unlocking the emotional world of visual media: An overview of the science, research, and impact of understanding emotion, Proceedings of the IEEE 111 (2023) 1236–1286.

[3] 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.

[4] H. Aviezer, Y. Trope, A. Todorov, Body cues, not facial expressions, discriminate between intense positive and negative emotions, Science 338 (2012) 1225–1229.

[5] P. Ekman, Darwin, deception, and facial expression, Annals of the new York Academy of sciences 1000 (2003) 205–221.

[6] Z. Yu, B. Zhou, J. Wan, P. Wang, H. Chen, X. Liu, S. Z. Li, G. Zhao, Searching multi-rate and multi-modal temporal enhanced networks for gesture recognition, IEEE Transactions on Image Processing 30 (2021) 5626–5640.

[7] W. Peng, X. Hong, H. Chen, G. Zhao, Learning graph convolutional network for skeleton-based human action recognition by neural searching, in: Proceedings of the AAAI conference on artificial intelligence, volume 34, 2020, pp. 2669–2676.

[8] D. Guo, K. Li, B. Hu, Y. Zhang, M. Wang, Benchmarking micro-action recognition: Dataset, method, and application, IEEE Transactions on Circuits and Systems for Video Technology (2024).

[9] D. Guo, X. Li, K. Li, H. Chen, J. Hu, G. Zhao, Y. Yang, M. Wang, Mac 2024: Micro-action analysis grand challenge, in: Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 11304–11305.

[10] A. Shah, H. Chen, G. Zhao, Representation learning for topology-adaptive micro-gesture recognition and analysis, in: IJCAI-MIGA Workshop & Challenge on Micro-gesture Analysis for Hidden Emotion Understanding (MiGA) July 21, 2023 Macao, China, Redaktion Sun SITE, 2023.

[11] A. Shah, H. Chen, G. Zhao, Naive data augmentation might be toxic: Data-prior guided self-supervised representation learning for micro-gesture recognition, in: 2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG), IEEE, 2024, pp. 1–9.

[12] 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.

[13] 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, in: 2019 14th

IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), IEEE, 2019, pp. 1–8.

[14] 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 (CVPR), 2022, pp. 2969–2978.

[15] G. Chen, F. Wang, K. Li, Z. Wu, H. Fan, Y. Yang, M. Wang, D. Guo, Prototype learning for micro-gesture classification, in: MiGA@ IJCAI, 2024.

[16] H. Zhou, Q. Liu, Y. Wang, Learning discriminative representations for skeleton based action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 10608–10617.

[17] H. Huang, Y. Wang, L. Kerui, Z. Xia, Multi-modal micro-gesture classification via multi-scale heterogeneous ensemble network, in: MiGA@ IJCAI, 2024.

[18] H. Huang, X. Guo, W. Peng, Z. Xia, Micro-gesture classification based on ensemble hypergraph-convolution transformer., in: MiGA@ IJCAI, 2023.

[19] X. Huang, H. Zhou, K. Yao, K. Han, Froster: Frozen clip is a strong teacher for open-vocabulary action recognition, ICLR2024 (2024).

[20] H. Chen, X. Liu, J. Shi, G. Zhao, Temporal hierarchical dictionary guided decoding for online gesture segmentation and recognition, IEEE Transactions on Image Processing 29 (2020) 9689–9702.

[21] H. Chen, X. Liu, G. Zhao, Temporal hierarchical dictionary with hmm for fast gesture recognition, in: 2018 24th international conference on pattern recognition (ICPR), IEEE, 2018, pp. 3378–3383.

[22] Y. Wang, L. Kerui, H. Huang, Z. Xia, Micro-gesture online recognition with dual-stream multi-scale transformer in long videos, in: MiGA@ IJCAI, 2024.

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

[24] P. Liu, F. Wang, K. Li, G. Chen, Y. Wei, S. Tang, Z. Wu, D. Guo, Micro-gesture online recognition using learnable query points, in: MiGA@ IJCAI, 2024.

[25] K. Li, D. Guo, G. Chen, X. Peng, M. Wang, Joint skeletal and semantic embedding loss for micro-gesture classification, MiGA@ IJCAI (2023).

[26] K. Xia, L. Wei, L. Yu, A spatio-temporal event transformer on versatile tasks for human behavior analysis, in: MiGA@ IJCAI, 2024.