MM-Gesture: Towards Precise Micro-Gesture Recognition through Multimodal Fusion

Jihao Gu[1], Fei Wang[2,5,6], Kun Li[3,*], Yanyan Wei[2,4], Zhiliang Wu[3] and Dan Guo[2,5]

[1]*University College London (UCL), Gower Street, London, WC1E 6BT, UK*

[2]*School of Computer Science and Information Engineering, School of Artificial Intelligence, Hefei University of Technology (HFUT)*

[3]*ReLER, CCAI, Zhejiang University, China*

[4]*Key Laboratory of Knowledge Engineering with Big Data (HFUT), Ministry of Education*

[5]*Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, China*

[6]*Xinsight Lab, Research Institute, Hefei Zhongjuyuan Intelligent Technology Co., Ltd., China*

Abstract

In this paper, we present MM-Gesture, the solution developed by our team HFUT-VUT, which ranked 1st in the micro-gesture classification track of the 3rd MiGA Challenge at IJCAI 2025, achieving superior performance compared to previous state-of-the-art methods. MM-Gesture is a multimodal fusion framework designed specifically for recognizing subtle and short-duration micro-gestures (MGs), integrating complementary cues from joint, limb, RGB video, Taylor-series video, optical-flow video, and depth video modalities. Utilizing PoseConv3D and Video Swin Transformer architectures with a novel modality-weighted ensemble strategy, our method further enhances RGB modality performance through transfer learning pre-trained on the larger MA-52 dataset. Extensive experiments on the iMiGUE benchmark, including ablation studies across different modalities, validate the effectiveness of our proposed approach, achieving a top-1 accuracy of 73.213%. Code is available at: https://github.com/momiji-bit/MM-Gesture.

Keywords

Micro-Gesture, Action Recognition, Multi-modal, Ensemble Fusion, Transfer Learning

1. Introduction

Micro-Gestures (MGs) [1, 2, 3, 4], defined as spontaneous and fine-grained movements, such as nose touching, hair scratching, or subtle finger rubs, encode rich affective and cognitive cues that rarely surface in conventional action recognition benchmarks. Compared with conventional actions [5, 6, 7, 8], MGs are unintentional, short-duration, and confined to small body regions, which makes them extremely difficult to capture and classify.

Due to the subtle changes and short duration of MGs, relying solely on a single modality (*e.g.*, RGB [9, 10, 11], skeleton [12, 13]) often captures merely partial characteristics of MGs, thus failing to fully and thoroughly exploit the comprehensive information latent in available data. Despite significant advancements in previous studies on micro-gesture and micro-action recognition [11, 10, 14, 15, 16, 17, 18], most existing approaches remain confined to utilizing limited modalities, such as RGB combined with skeleton data [19, 20, 21]. However, these methods have not sufficiently leveraged the abundant and complementary information conveyed by multi-modal.

In this work, we propose a novel multi-modal fusion framework **MM-Gesture** tailored explicitly for the challenging task of MGs classification. Specifically, we construct baseline models leveraging PoseConv3D [21] and Video Swin Transformer [9, 6], integrating information across six complementary modalities: **joint, limb, RGB video, Taylor video, optical flow video, and depth video**. In addition,

to enhance the performance of the RGB modality, we apply transfer learning by pre-training on the Micro-Action 52 dataset [11] and fine-tuning on the iMiGUE dataset [2].

The key contributions of this paper can be summarized as follows:

- We present an integrated multi-modal MGs classification network that utilizes complementary information from six diverse modalities: joint, limb, RGB video, Taylor video, optical flow video, and depth video.
- We propose an effective ensemble fusion method capable of efficiently integrating six modalities, enabling the joint exploitation of modality-specific strengths for improved MGs classification accuracy.
- Extensive experiments on the iMiGUE dataset [2] demonstrate that the proposed MM-Gesture achieves state-of-the-art performance, reaching a Top-1 accuracy of 73.213%, which is the highest reported accuracy in previous Micro-gesture Analysis (MiGA) challenges.

2. Related Work

Micro-Gestures (MGs) are becoming increasingly important in understanding human emotions, focusing on subtle body movements in daily interactions. Advances in this field have been driven by the development of large benchmark datasets and sophisticated model architectures [1, 2, 3, 11]. Key datasets include the SMG dataset [3], which consists of recordings from 40 participants engaged in storytelling, capturing upper limb micro-gestures and emotional states. The iMiGUE dataset [2] offers identity-free videos of 72 athletes at press conferences, annotated with 32 micro-gesture categories for analyzing both actions and emotions. The MA-52 dataset [11] expands the focus to full-body micro-actions, with 22,000 samples covering 52 action-level and 7 body-level categories, sourced from psychological interviews to recognize subtle visual cues.

Current models primarily focus on limited modalities. RGB-based methods leverage spatial-temporal modeling strategies, such as a pure Transformer backbone with shifted 3D local attention windows [9]. MANet [11] integrates SE and TSM modules with semantic embedding loss for fine-grained micro-action recognition. Skeleton-based approaches include a 3D-CNN model with joint and semantic embedding losses [12], and an EHCT framework [13] employs hypergraph-based attention and ensemble Transformers [22, 23] to capture high-order joint relations and address class imbalance. In contrast, skeleton sequences can be encoded as 3D heatmaps and fused with RGB inputs through a dual-branch multimodal network [21]. Inspired by this network, Chen *et al.* [19] adopt channel-wise cross-attention and prototype refinement to enhance feature fusion and category discrimination, while Huang *et al.* [24] design a multi-scale heterogeneous fusion network. Recently, Li *et al.* [10] propose a hierarchical prototype-based calibration method to resolve ambiguity in fine-grained actions. Overall, current methods only focus on the RGB or skeleton data.

To exploit the complementarity between different multimodal data, we propose the MM-Gesture model, adopting a comprehensive multimodal approach that integrates six modalities: joint, limb, RGB video, Taylor video, optical flow video, and depth video. This approach enables a deeper understanding and representation of micro-gestures, capturing their nuances and dynamics. Additionally, we leverage transfer learning from the MA-52 dataset to infuse valuable prior knowledge into the RGB modality, further enhancing its recognition accuracy. Consequently, our model improves performance on existing benchmarks and paves the way for advanced applications in human emotion understanding through micro-gesture analysis.

3. Methodology

3.1. Data Pre-processing

We adopted the RGB videos ($\mathbf{R} \in \mathbb{R}^{T \times H \times W \times 3}$) provided by the official dataset, along with a subset of 36 skeleton keypoints (V) selected from the original 137 points, to form the input joint data ($\mathbf{J} \in \mathbb{R}^{T \times V \times 2}$).
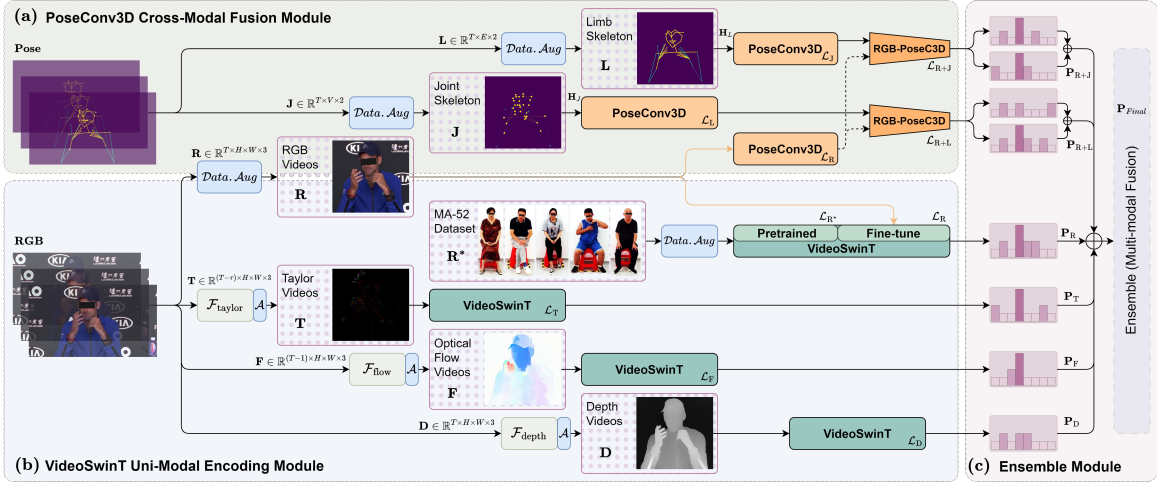
Figure 1: Pipeline of the proposed multimodal micro-gesture recognition framework (**MM-Gesture**), which consists of three key components: (a) Cross-Modal Fusion Module, (b) Uni-modal Embedding Module, and (c) Ensemble Module.

These cleaned keypoints focus specifically on the upper body, hands, and facial joints. Additionally, we constructed input limb data ($\mathbf{L} \in \mathbb{R}^{T \times E \times 2}$) by computing spatial differences between adjacent joint pairs defined by the skeletal edges (E) connecting the selected keypoints.

To effectively capture multi-modal gesture information, we employ advanced, off-the-shelf modality extraction methods to generate complementary auxiliary modalities. Specifically, we utilize Taylor-series temporal expansion videos, optical-flow videos, and depth-estimation videos, each modality providing distinct yet complementary gesture-related information. By leveraging the ensemble among these diverse modalities, our proposed MM-Gesture model effectively exploits multi-modal feature complementarity.

$$
\begin{aligned}
\mathbf{T} \in \mathbb{R}^{(T-\tau) \times H \times W \times 3}, \quad &\mathbf{F} \in \mathbb{R}^{(T-1) \times H \times W \times 3}, \quad \mathbf{D} \in \mathbb{R}^{T \times H \times W \times 3}, \\
\mathbf{T}_t = \mathscr{F}_{\text{taylor}}(\mathbf{R}_{t:t+\tau}), \quad &\mathbf{F}_t = \mathscr{F}_{\text{flow}}(\mathbf{R}_{t:t+1}), \quad \mathbf{D}_t = \mathscr{F}_{\text{depth}}(\mathbf{R}_t),
\end{aligned}
\tag{1}
$$

where each symbol is defined as follows:

- T: Temporal length of the input RGB video.
- H, W: Height and Width of the input RGB video frames.
- τ: Temporal window length for computing the truncated Taylor-series expansion.
- $\mathbf{R}_t$: The RGB frame at time step t.
- $\mathscr{F}_{taylor}$: The Taylor-series-based video calculated according to the approach [25], where K denotes the maximum order of the truncated Taylor-series expansion and τ represents the temporal window length used for aggregating local temporal context.
- $\mathscr{F}_{flow}$: The optical-flow-based modality computed using the MemFlow network [26], which estimates optical flow representations $\mathbf{F}_t$ from consecutive frames $\mathbf{R}_t$ and $\mathbf{R}_{t+1}$.
- $\mathscr{F}_{depth}$: The depth-estimation-based modality generated using the monocular depth estimation algorithm [27], resulting in depth representations $\mathbf{D}_t$.

3.2. Network Architecture

As shown in Figure 1, the proposed multi-modal micro-gesture recognition framework (**MM-Gesture**) consists of three main modules:

Cross-Modal Fusion Module: In this module, skeletal coordinates are initially transformed into Gaussian heatmap-based 3D volumes (**H**) for Joint and Limb modalities individually. RGB, Joint, and

Limb modalities are all separately trained through PoseConv3D [21], capturing spatial-temporal skeleton dynamics and RGB spatial context, respectively. Subsequently, the extracted RGB and skeleton features are combined via a cross-modal fusion training stage to exploit complementary information between these modalities comprehensively.

Uni-Modal Encoding Module: We leverage the VideoSwinT network [9] to independently encode four distinct modalities: RGB frames, Taylor-based temporal encoding, optical flow (computed via MemFlow), and depth estimates. Specifically, for the RGB modality, we first employ transfer learning by pretraining VideoSwinT on the MA-52 dataset and subsequently fine-tune the pretrained model on the iMiGUE dataset. For the remaining modalities (Taylor, optical flow, and depth), VideoSwinT is directly trained from scratch on the iMiGUE dataset. VideoSwinT uses a 3D shifted-window self-attention mechanism that effectively captures fine-grained spatial-temporal details within each modality.

Ensemble Module: Probabilities from the PoseConv3D Cross-Modal Fusion Module and VideoSwinT Uni-Modal Encoding Module are combined via weighted ensemble, with weights set empirically according to validation performance. This integration approach effectively exploits modality complementarity, improving robustness and accuracy in micro-gesture recognition.

3.3. PoseConv3D Cross-Modal Fusion Module

To effectively align skeleton-based information (consisting of joints and limbs) with RGB video representations and facilitate fine-grained complementary interactions across these modalities, we adopt PoseConv3D [21] for cross-modal integration.

Specifically, we first transform the 2D coordinates of skeletal keypoints into heatmap-based representations. By applying Gaussian distributions and calculating the heatmap values using the point-to-segment distance formula, we compute and stack the heatmaps of each keypoint across all frames to generate 3D heatmap volumes. The resulting heatmaps are as follows:

$$\mathbf{H}_J \in \mathbb{R}^{T \times H \times W \times V}, \quad \mathbf{H}_L \in \mathbb{R}^{T \times H \times W \times E}, \tag{2}$$

where $\mathbf{H}_J$ denotes the joint-position heatmaps, and $\mathbf{H}_L$ denotes the limb-connection heatmaps. Here, T is the total number of frames, V is the number of skeletal joints, and E is the number of skeletal limbs (connections between joints). H, and W represent the spatial resolution (height and width) of each heatmap. Subsequently, the RGB frames $\mathbf{R} \in \mathbb{R}^{T \times H \times W \times 3}$ and skeleton heatmaps $\mathbf{H}_J, \mathbf{H}_L$ are taken as input data.

Prior to network training, data augmentation processes (*e.g.*, scaling, cropping) are consistently applied to both RGB video frames and skeleton heatmap modalities to enhance data diversity and improve model robustness. Subsequently, the augmented data from each modality is separately forwarded into the PoseConv3D module, which extracts deep spatiotemporal feature representations. The PoseConv3D network generates modality-specific predictions denoted formally as $\hat{\mathbf{y}}_m$, where $m \in \{R, J, L\}$ indicates RGB, joint heatmap, and limb heatmap modalities, respectively. Each modality-specific network is initially pretrained independently by minimizing the cross-entropy (CE) classification loss:

$$\mathcal{L}_m = \text{CE}\left(\hat{\mathbf{y}}_m, y\right), \quad m \in \{R, J, L\}, \tag{3}$$

where y denotes the ground-truth action labels.

Next, we conduct a joint fine-tuning procedure by simultaneously optimizing combined RGB and skeleton-based modalities using the following paired-training losses:

$$\mathcal{L}_{R+J} = \mathcal{L}_R + \mathcal{L}_J, \quad \mathcal{L}_{R+L} = \mathcal{L}_R + \mathcal{L}_L. \tag{4}$$

During model inference, the predictions yielded by distinct modalities are integrated at the probability level via a late fusion strategy. Formally, let $P^\star = \text{SoftMax}(\hat{\mathbf{y}}^\star)$, $\star \in \{R, J, L\}$, represent modality-specific probability distributions. We then fuse predictions through average fusion to achieve final predictive distributions:

$$\mathbf{P}_{R+J} = \frac{1}{2}(\mathbf{P}_R + \mathbf{P}_J), \quad \mathbf{P}_{R+L} = \frac{1}{2}(\mathbf{P}_R + \mathbf{P}_L). \tag{5}$$

3.4. VideoSwinT Uni-Modal Encoding Module

Unlike existing skeleton-video modality fusion methods, we propose a multimodal framework based on the VideoSwinT [9], which encodes RGB video, optical flow video, Taylor-expanded video, and depth video. This encoding strategy effectively integrates color, texture, dynamic motion, and geometric structural information to better capture multidimensional micro-action features, thus enabling more fine-grained action recognition.

Specifically, we independently optimize each modality-specific backbone by minimizing the cross-entropy (CE) classification loss. Prior to training on the target iMiGUE dataset, the RGB modality network is initially pretrained on the MA-52 dataset ($\mathbf{R}^{\star} \in \mathbb{R}^{T \times H \times W \times 3}$) [11], which provides extensive coverage of 52 types of micro-actions. After pretraining, the RGB modality network is fine-tuned on the iMiGUE dataset along with other modalities. The loss functions for pretraining and fine-tuning, along with the probability computation, are formulated as follows:

$$
\begin{aligned}
\mathscr{L}_m &= \text{CE}(\hat{\mathbf{y}}_m, y), \quad m \in \{\text{R}^{\star}, \text{R}, \text{T}, \text{F}, \text{D}\}, \\
\mathbf{P}_m &= \text{SoftMax}(\hat{\mathbf{y}}_m), \quad m \in \{\text{R}, \text{T}, \text{F}, \text{D}\}.
\end{aligned}
\tag{6}
$$

3.5. Ensemble Module

In the final ensemble stage, we introduce a probability-based weighted fusion strategy to effectively aggregate predictions derived from multiple modality-specific networks. Specifically, class probability vectors independently output by the PoseConv3D (**RGB + J**, **RGB+L**) and VideoSwin Transformer (**RGB***, **Taylor**, **Flow**, **Depth**) models are integrated using empirically determined weights obtained via validation-set performance.

The ensemble prediction ($\mathbf{P}_{\text{final}} \in \mathbb{R}^{cls}$) is computed by summing the weighted contributions of individual modality-specific probabilities, as follows:

$$
\mathbf{P}_{\text{final}} = \sum w_i \mathbf{P}_i, \quad i \in \{\text{R+J}, \text{R+L}, \text{R}, \text{T}, \text{F}, \text{D}\}
\tag{7}
$$

where each weight w_i is selected based on the classification performance observed on validation samples.

This proposed ensemble-based fusion mechanism enables comprehensive exploitation of the complementary strengths inherent in multiple modality-specific models, thereby significantly improving the robustness and overall effectiveness of our multi-modal micro-gesture recognition framework.

4. Experiments

4.1. Experimental Setup

Dataset. iMiGUE (identity-free video dataset for Micro-Gesture Understanding and Emotion analysis) dataset [2] consists of micro-gestures (MGs) primarily involving upper limbs, collected from post-match press conference videos of professional tennis players. It includes 31 MG categories and an additional non-MG class, comprising a total of 18,499 labeled MG samples annotated from 359 long video sequences (ranging from 0.5 to 26 minutes), totaling approximately 3.77 million frames. The dataset provides two modalities: RGB videos and corresponding 2D skeletal joint data extracted via OpenPose. iMiGUE adopts a cross-subject evaluation protocol, splitting 72 subjects into 37 for training and 35 for testing, with 12,893 samples in the training set, 777 in validation, and 4,562 in testing. In addition, we pretrain the proposed method on the **Micro-Action 52** [11] dataset and then fine-tune it on the iMiGUE dataset. Micro-Action 52 is a large-scale, whole-body micro-action dataset collected by a professional interviewer to capture unconscious human micro-action behaviors. The dataset contains 22,422 (22.4K) samples interviewed from 205 participants, where the annotations are categorized into two levels: 7 *body-level* and 52 *action-level* micro-action categories. There are 11,250, 5,586, and 5,586 instances in the training, validation, and test sets, respectively.

Table 1

Top-3 micro-gesture classification results from MiGA Challenges (2023–2025). Results are sourced from official competition leaderboards[1][2][3]. **J** denotes the Joint modality; **L** denotes the Limb modality; **R** denotes the RGB video modality; **T** denotes the Taylor video modality; **F** denotes the Optical Flow video modality; **D** denotes the Depth video modality.

Rank	Team	Backbone	Modality	Top-1 Acc (%)
MiGA'25 1st	gkdx2 (**Ours**)	PoseConv3D+VideoSwinT	**J + L + R + T + F + D**	**73.213**
MiGA'25 2nd	awuniverse	-	-	68.697
MiGA'25 3rd	Lonelysheep	PoseConv3D	**J + L**	67.010
MiGA'24 1st	HFUT-VUT [19]	PoseConv3D	**J + L + R**	70.254
MiGA'24 2nd	NPU-MUCIS [20]	Res2Net3D+GCN	**J + R**	70.188
MiGA'24 3rd	ywww11	PoseConv3D+CLIP	**J + R**	68.917
MiGA'23 1st	HFUT-VUT [12]	PoseConv3D	**J + L**	64.12
MiGA'23 2nd	NPU-Stanford [13]	Hyperformer	**J**	63.02
MiGA'23 3rd	ChenxiCui [28]	-	-	62.63

Evaluation Metrics. For the micro-gesture classification challenge, we employ top-1 accuracy as the evaluation metric to quantitatively assess classification performance.

Implementation Details. The provided dataset includes original RGB videos and skeletal data extracted using OpenPose, featuring 137 full-body keypoints. To optimize data, we select 36 keypoints for the upper-body, facial landmarks, and hands. We also enhance data representation by generating additional modalities: depth using the method by *Chen et al.* [27], Taylor video modality via *Wang et al.*'s [25] approximation, and optical flow through *Dong et al.*'s [26] MemFlow approach. For modeling, PoseConv3D [21] is used to capture spatial-temporal dynamics in skeletal information (J), limb connections (L), and combined RGB with skeletal data (RGB+J and RGB+L). VideoSwin Transformer [9] is applied to RGB, depth, Taylor, and optical flow modalities for spatial-temporal processing. To enhance robustness, we perform transfer learning with VideoSwinT: initially pretraining on RGB data from Micro-Action 52 (MA-52) [11], followed by fine-tuning on the iMiGUE dataset [2]. Finally, we employ an ensemble fusion strategy, assigning weights to each modality based on contribution and correlation. We integrate **RGB***, **Taylor**, **Flow**, and **Depth** from VideoSwin, along with **RGB+Joint** and **RGB+Limb** from PoseConv3D.

4.2. Experimental Results

We evaluated the proposed method on the iMiGUE dataset and compared its performance against state-of-the-art methods reported in the MiGA Challenges from 2023 to 2025. As presented in Table 1, we provide the classification results of the top three competitors from these three consecutive editions, clearly demonstrating the consistent superiority of our proposed method over previous best-performing approaches across all years. Specifically, our approach achieved a Top-1 accuracy of 73.213%, ranking first in the 2025 competition, significantly outperforming the second-place accuracy of 68.697%. Compared with the best performance in the 2024 MiGA Challenge, our method realized an improvement of approximately 3%, thus substantially exceeding the results from the 2023 edition as well.

Here, we conduct comprehensive experimental settings to evaluate multiple modalities, including skeleton data (joints and limbs), RGB frames, Taylor series approximation videos (Taylor), optical flow, and depth information. As shown in Table 2, two backbone frameworks, namely PoseConv3D [21] and VideoSwin [9], were employed to thoroughly explore performance across various modality combinations. Experimental outcomes demonstrate that while single-modality inputs generally show

Table 2
Comparison of classification performance using different combinations of modalities and backbone architectures on the iMiGUE test set. The evaluated backbone models include PoseConv3D [21] and VideoSwinT [9]. We evaluate six modalities: Joint, Limb, RGB, Taylor, Optical Flow, and Depth. Particularly, RGB* denotes that transfer learning was adopted by first pre-training on the Micro-Action 52 dataset [11] and subsequently fine-tuning on the iMiGUE dataset [2].

Backbone	Joint	Limb	RGB	RGB*	Taylor	Flow	Depth	Top-1 Acc (%)
PoseConv3D	✓							65.256
PoseConv3D		✓						64.686
PoseConv3D			✓					64.511
PoseConv3D	✓	✓						67.229
PoseConv3D	✓		✓					68.917
PoseConv3D		✓	✓					68.917
VideoSwinT			✓					65.629
VideoSwinT				✓				66.615
VideoSwinT					✓			62.845
VideoSwinT						✓		61.617
VideoSwinT							✓	65.212
PoseConv3D+VideoSwinT	✓			✓				70.955
PoseConv3D+VideoSwinT		✓		✓				70.802
PoseConv3D+VideoSwinT	✓	✓		✓				71.416
PoseConv3D+VideoSwinT	✓	✓		✓	✓			72.095
PoseConv3D+VideoSwinT	✓	✓		✓	✓	✓		72.227
PoseConv3D+VideoSwinT	✓	✓		✓	✓	✓	✓	72.644
MM-Gesture (**Ours**)	✓	✓	✓	✓	✓	✓	✓	**73.213**

moderate competitiveness, they nevertheless yield relatively lower accuracies, highlighting the inherent challenges of relying on a single modality in micro-gesture classification tasks. However, the incorporation of multiple modalities consistently results in enhanced performance, clearly emphasizing the complementary and distinctive nature of the various modalities in improving classification accuracy.

Our subsequent multimodal fusion experiments verify the complementary nature of diverse data streams. Specifically, integrating skeleton (joint and limb) data with RGB frames results in an accuracy improvement to 71.416%, clearly demonstrating the strength of combining structural and appearance-based representations. Incorporating the Taylor modality further boosts accuracy to 72.096%, reflecting benefits from pixel-level temporal-spatial approximations that effectively capture subtle dynamic gestures. Additional integration of optical flow and depth modalities improves performance even further, reaching an accuracy of 72.644%, confirming their roles as valuable supplementary information sources. Ultimately, through an optimized multimodal fusion weighting strategy, our method achieves a Top-1 accuracy of 73.213%. These results strongly affirm the advantages of properly designed multimodal fusion techniques and emphasize the efficacy and robustness of the presented approach over previously published state-of-the-art methods in micro-gesture recognition tasks.

5. Conclusion

In this paper, we proposed **MM-Gesture**, a novel multimodal ensemble framework for micro-gesture recognition. Our method integrates complementary features from six modalities—skeleton, limb, RGB, Taylor series approximation, optical flow, and depth—to leverage their distinct fine-grained characteristics. Additionally, we employed transfer learning by pretraining the RGB-based model on the Micro-Action 52 dataset before fine-tuning on the target iMiGUE dataset. Experiments demonstrate that our multimodal fusion significantly outperforms single or fewer modality baselines. Our model achieved a top-1 accuracy of 73.213% on the challenging iMiGUE dataset, ranking first in the 3rd MiGA Competition at IJCAI 2025.

For future work, we aim to explore the integration of Multimodal large language models (MLLMs) [29, 30] and skeleton-based micro-gesture encoders. We plan to utilize MLLMs' rich semantic understanding and extensive prior knowledge to enhance micro-gesture recognition through interactive prompts and contextual reasoning, further advancing multimodal and affective human behavior understanding. Additionally, we will incorporate modalities such as gaze [31], audio [32], and remote photoplethysmography (rPPG) [33] to enable comprehensive multimodal emotion analysis.

Acknowledgments

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT solely for grammar and spelling checks and minor language refinement. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

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, in: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition, 2019, pp. 1–8.

[2] 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.

[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. Chen, B. W. Schuller, E. Adeli, G. Zhao, The 2nd challenge on micro-gesture analysis for hidden emotion understanding (miga) 2024: Dataset and results, in: MiGA 2024: Proceedings of IJCAI 2024 Workshop&Challenge on Micro-gesture Analysis for Hidden Emotion Understanding (MiGA 2024) co-located with 33rd International Joint Conference on Artificial Intelligence (IJCAI 2024), 2024.

[5] K. Li, D. Guo, M. Wang, Proposal-free video grounding with contextual pyramid network, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2021, pp. 1902–1910.

[6] F. Wang, K. Li, Y. Nie, Z. Duan, P. Zou, Z. Wu, Y. Wang, Y. Wei, Exploiting ensemble learning for cross-view isolated sign language recognition, arXiv preprint arXiv:2502.02196 (2025).

[7] M. Balazia, P. Müller, Á. L. Tánczos, A. v. Liechtenstein, F. Bremond, Bodily behaviors in social interaction: Novel annotations and state-of-the-art evaluation, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 70–79.

[8] D. Li, B. Xing, X. Liu, B. Xia, B. Wen, H. Kälviäinen, Deemo: De-identity multimodal emotion recognition and reasoning, arXiv preprint arXiv:2504.19549 (2025).

[9] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, H. Hu, Video swin transformer, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 3202–3211.

[10] K. Li, D. Guo, G. Chen, C. Fan, J. Xu, Z. Wu, H. Fan, M. Wang, Prototypical calibrating ambiguous samples for micro-action recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 39, 2025, pp. 4815–4823.

[11] D. Guo, K. Li, B. Hu, Y. Zhang, M. Wang, Benchmarking micro-action recognition: Dataset, methods, and applications, IEEE Transactions on Circuits and Systems for Video Technology 34 (2024) 6238–6252.

[12] K. Li, D. Guo, G. Chen, X. Peng, M. Wang, Joint skeletal and semantic embedding loss for micro-gesture classification, arXiv preprint arXiv:2307.10624 (2023).

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

[14] K. Li, P. Liu, D. Guo, F. Wang, Z. Wu, H. Fan, M. Wang, Mmad: Multi-label micro-action detection in videos, arXiv preprint arXiv:2407.05311 (2024).

[15] K. Li, D. Guo, G. Chen, F. Liu, M. Wang, Data augmentation for human behavior analysis in multi-person conversations, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 9516–9520.

[16] J. Gu, K. Li, F. Wang, Y. Wei, Z. Wu, H. Fan, M. Wang, Motion matters: Motion-guided modulation network for skeleton-based micro-action recognition, in: Proceedings of the 33rd ACM International Conference on Multimedia, 2025.

[17] S. Sun, D. Liu, J. Dong, X. Qu, J. Gao, X. Yang, X. Wang, M. Wang, Unified multi-modal unsupervised representation learning for skeleton-based action understanding, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 2973–2984.

[18] J. Dong, S. Sun, Z. Liu, S. Chen, B. Liu, X. Wang, Hierarchical contrast for unsupervised skeleton-based action representation learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 37, 2023, pp. 525–533.

[19] G. Chen, F. Wang, K. Li, Z. Wu, H. Fan, Y. Yang, M. Wang, D. Guo, Prototype learning for micro-gesture classification, arXiv preprint arXiv:2408.03097 (2024).

[20] H. Huang, Y. Wang, L. Kerui, Z. Xia, Multi-modal micro-gesture classification via multiscale heterogeneous ensemble network, MiGA@ IJCAI (2024).

[21] 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, 2022, pp. 2969–2978.

[22] F. Wang, D. Guo, K. Li, M. Wang, Eulermormer: Robust eulerian motion magnification via dynamic filtering within transformer, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, 2024, pp. 5345–5353.

[23] F. Wang, D. Guo, K. Li, Z. Zhong, M. Wang, Frequency decoupling for motion magnification via multi-level isomorphic architecture, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 18984–18994.

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

[25] L. Wang, X. Yuan, T. Gedeon, L. Zheng, Taylor videos for action recognition, in: Forty-first International Conference on Machine Learning, 2024.

[26] Q. Dong, Y. Fu, Memflow: Optical flow estimation and prediction with memory, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 19068–19078.

[27] S. Chen, H. Guo, S. Zhu, F. Zhang, Z. Huang, J. Feng, B. Kang, Video depth anything: Consistent depth estimation for super-long videos, arXiv preprint arXiv:2501.12375 (2025).

[28] H. Xu, L. Cheng, Y. Wang, S. Tang, Z. Zhong, Towards fine-grained emotion understanding via skeleton-based micro-gesture recognition, arXiv preprint arXiv:2506.12848 (2025).

[29] Y. Xu, L. Zhu, Y. Yang, Mc-bench: A benchmark for multi-context visual grounding in the era of mllms, arXiv preprint arXiv:2410.12332 (2024).

[30] Y. Xu, L. Zhu, Y. Yang, Gg-editor: Locally editing 3d avatars with multimodal large language model guidance, in: Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 10910–10919.

[31] F. Liu, K. Li, Z. Zhong, W. Jia, B. Hu, X. Yang, M. Wang, D. Guo, Depth matters: Spatial proximity-based gaze cone generation for gaze following in wild, ACM Transactions on Multimedia Computing, Communications and Applications 20 (2024) 1–24.

[32] J. Zhao, F. Wang, K. Li, Y. Wei, S. Tang, S. Zhao, X. Sun, Temporal-frequency state space duality: An efficient paradigm for speech emotion recognition, in: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing, 2025, pp. 1–5.

[33] W. Qian, K. Li, D. Guo, B. Hu, M. Wang, Cluster-phys: Facial clues clustering towards efficient remote physiological measurement, in: Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 330–339.