Lappeenranta-Lahti University of Technology LUT
School of Engineering Science
Computational Engineering and Technical Physics
Computer Vision and Pattern Recognition

Jussi Juvonen

Student No: 0575457
Course: Machine Vision and Digital Image Analysis BM40A0801
Date: March 29, 2025

IMAGE-LEVEL MICRO GESTURE CLASSIFICATION

ABSTRACT

Lappeenranta-Lahti University of Technology LUT
School of Engineering Science
Computational Engineering and Technical Physics
Computer Vision and Pattern Recognition

Jussi Juvonen

IMAGE-LEVEL MICRO GESTURE CLASSIFICATION

2025

12 pages.

Keywords: multimodal fusion, micro-gesture recognition, spatial-temporal graph convolutional networks, skeletal-based action recognition, deep residual learning, computer vision

An intelligent multimodal system for micro-gesture recognition is presented. The study focuses on the integration of RGB video analysis and skeleton-based motion understanding using a two-stream fusion model. A spatial-temporal graph convolutional network is employed to process skeletal data, while deep residual learning is utilized to extract visual features from RGB frames. The system was evaluated on a dataset partitioned into training, validation, and testing subsets. An ablation study was conducted to assess the contributions of individual modalities and their fusion. Results indicate that the fusion model achieved the highest accuracy, albeit with only marginal improvement over the RGB-only configuration. The skeleton model exhibited challenges due to extraction errors and class imbalance. Architectural limitations in the fusion strategy were indentified as potential areas for improvement. It is concluded that further refinements in cross-modal interaction and processing methods are required to enhance performance and approach state-of-the-art benchmarks.

CONTENTS

1 INTRODUCTION

Micro gestures are subtle and brief movements that can convey emotional and communicative cues. The ability to recognize non-verbal signals is important to improve human-computer interaction and developing more intuitive user interfaces.

This project uses the iMiGUE dataset, which contains approximately 18000 annotated clips across 32 micro-gesture categories. The primary task is to classify these micro gestures accurately. By establishing a baseline classification model.

The original research introducing the iMiGUE dataset established baseline performance using a range of methodologies, including Recurrent Neural Networks (RNNs), Graph Convolutional Networks (GCNs), and Convolutional Neural Networks (CNNs). Notably a Temporal Shift Module (TSM) model utilizing RGB data achieved a Top-1 accuracy of 61.10%. [1]

The iMiGUE dataset poses a challenge due to severe class imbalance. For example, while one class may be represented by only a single video clip, another class might contain several thousand clips. This imbalance can significantly affect the performance of a classification model, potentially leading it to favor dominant classes over the underrepresented ones.

The detailed nature of micro-gesture categories means that an effective model needs to recognize small visual and movement differences. It could benefit from using integrating multiple data modalities, such as combining skeleton data with RGB data, to create a fusion model that captures complementary information.

2 LITERATURE REVIEW

To effectively process skeleton data within a multimodal framework for micro-gesture recognition, Spatial-Temporal Graph Convolutional Networks (STGCNs) offer a strong approach. STGCNs are specifically designed for action recognition using skeleton sequences. These networks represent the human skeleton as a graph, where joints correspond to nodes and natural connections between adjacent joints form the edges. This graph-based representation allows the model to learn both the spatial configuration of the human body at any given time and the temporal evolution of these configurations over a sequence of frames [2].

Processing visual information from RGB images has greatly benefited from deep residual learning. In these networks, layers are designed to learn residual functions relative to their inputs, which simplifies the training of very deep architectures. Models like ResNet18 and ResNet50, for example, leverage this principle to build hierarchical representations of visual features. This approach not only eases the optimization process but has also has led to state-of-the-art performance across a range of image and video analysis tasks. [3]

Recent advancements in GCN-based models for skeleton-based action recognition have addressed several key limitations in traditional STGCNs. One such improvement is the introduction of adaptive graph convolutional networks, which allow for a more flexible approach to constructing skeletal graph structures. These adaptive methods dynamically adjust the graph based on the input data, leading to improved feature extraction and recognition accuracy. Additionally, multi-stream frameworks that incorporate bones, joints, and motion information combined with spatial, temporal, and channel attention mechanisms have been shown to enhance performance in action recognition tasks. The Extended Multi-stream Temporal-attention Adaptive GCN (EMS-TAGCN) is one such model that integrates these enhancements, achieving state-of-the-art results on benchmark datasets. [4]

3 METHOD DESCRIPTION

My solution combines RGB video analysis with skeleton-based motion understanding through a two-stream fusion model, implemented in PyTorch. This architecture was chosen to address the subtle spatial-temporal patterns in micro-gestures while mitigating class imbalance challenges in the iMiGUE dataset.

3.1 Multimodal architecture

The model consists of two parallel branches: RGB and Skeleton. The RGB branch utilizes ResNet18 pretrained on ImageNet as a spatial feature extractor. It processes frames independently, then applies temporal attention. The skeleton branch uses the OpenPose model (Body_25) to extract human body keypoints. Each skeleton frame consists of 25 keypoints, where each keypoint has (x, y) coordinates and a confidence score. These features are processed using STGCN. The skeleton branch uses STGCN, which captures joint relationships through three STGCN blocks. Finally, branches are fused by concatenating RGB and skeleton features. This is processed through two fully-connected layers (640→512→32) dimensions with ReLU activation and 30% dropout before final classification. The skeleton branch implementation was influenced by the STGCN framework described in [5]. An overview of the model is shown in Figure 1.

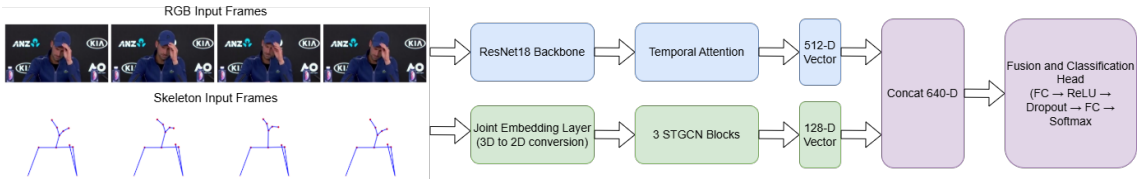

Figure 1. Overview of the multimodal two-stream fusion model for micro-gesture classification.

3.2 Implementation details

The model uses AdamW optimizer with weight decay of $5e-4$, with cosine annealing learning rate scheduling (Initial learning rate is $3e-4$). Cross-entropy loss was used.

RGB frames, originally captured at a 320×180 resolution undergo spatial normalization through a two-stage process. During training, randomly resized crops extract 224×224

patches from 256-pixel scaled images. For validation/testing, center crops provide consistent spatial references. All frames undergo ImageNet-standard normalization with mean $[0.485, 0.456, 0.406]$ and standard deviation $[0.229, 0.224, 0.225]$.

Skeleton data from OpenPose's body_25 model are processed at 320×176 resolution to maintain 16:9 aspect ratio. The keypoints are filtered with confidence based metrics. Keypoints with detection confidence below 0.1 are treated as missing values through a multi-stage process. First, valid joint positions (confidence ≥ 0.1) are normalized using per-coordinate statistics $(\mu_{\text{train}}, \sigma_{\text{train}}, 0.1)$ computed from the training set. Frames with completely invalid skeletons (all confidence scores < 0.1) receive zero-valued confidence channels, while partially valid frames retain only high-confidence joints through positional masking. Temporal processing standardizes variable-length sequences to four frames, through either random selection for clips with more than four frames, or cyclical repetition of existing frames.

4 EXPERIMENTS & RESULTS

The provided dataset was originally organized as a single training folder. For the experiments I partitioned this dataset into three subsets with a 60/20/20 split for training, validation, and testing, respectively. This partitioning was performed at the clip level to maintain temporal continuity.

In the training dataset, out of 7732 clips, 14 frames were skipped and no clip contained exclusively invalid frames. In the validation dataset, out of 2586 clips, 5 frames were skipped, and there was 1 clip where all frames were invalid. In the testing dataset, out of 2593 clips, 1 frame was skipped and no clip contained exclusively invalid frames.

To assess the contribution of each modality, an ablation study was conducted, comparing three configurations: the RGB stream only, the skeleton stream only, and the fusion model that combines both modalities. This study was designed to determine the individual performance of each modality as well as the benefit of combining them. All models were trained for 20 epochs using the same hyperparameters, optimizer, and scheduler. The table (1) summarizes the experimental results in terms of key metrics such as top-1, top-5 accuracy, and time elapsed:

Table 1. Ablation study results for individual modalities and the fusion model. Reported accuracies are from the test dataset. Reported times represent average time per epoch.

Modality	**Top-1 (%)**	**Top-5 (%)**	**Training Time (s)**	**Evaluation Time (s)**
RGB Only	53.9	90.6	146.1	25.9
Skeleton Only	33.7	71.2	**17.3**	**4.3**
Fusion Model	**55.2**	**91.2**	184.3	34.5

The fusion model achieved the best accuracy results, but only marginally outperformed the RGB-only baseline. The skeleton-only configuration showed a pronounced bias toward majority classes, achieving approximately 13.3% lower Top-1 accuracy than reported for STGCN architectures on comparable benchmarks [1]. Computational efficiency varied significantly, with skeleton processing requiring much lower computation cost. However this does not account for the time required to extract skeletons from the original images. Figure 2 shows the Top-1 accuracy plots from the ablation study and Figure 3 presents the loss plots. Figure 4 shows the confusion matrix of the fusion model on the testing dataset.

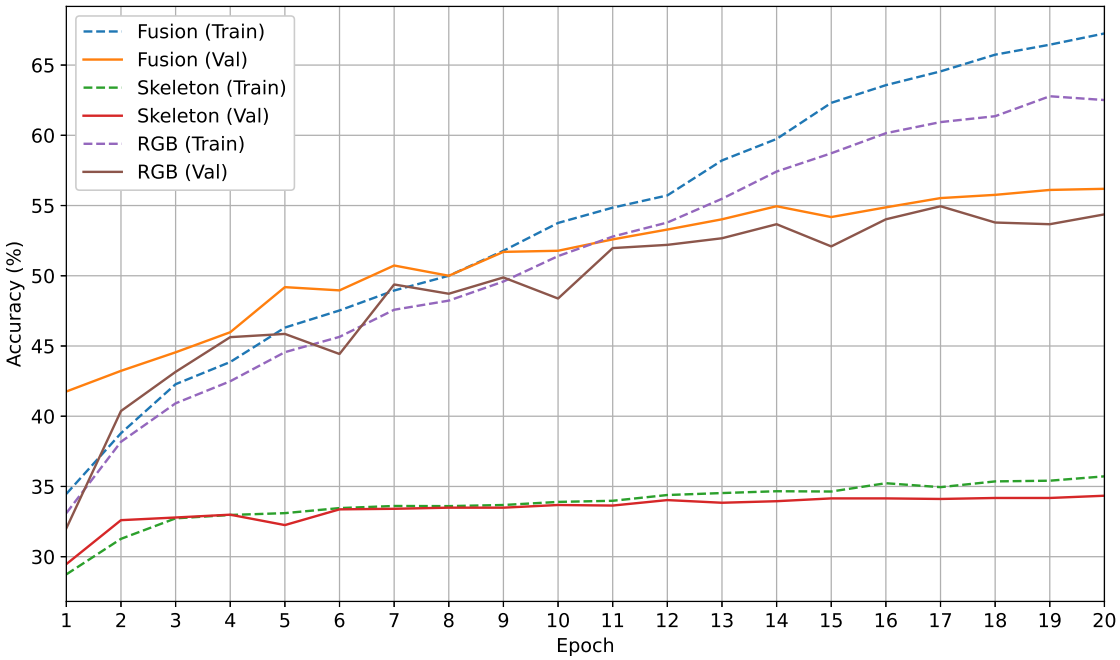

Figure 2. Combined top-1 accuracy plots from ablation study. Dashed lines indicate training data, while solid lines indicate validation data.

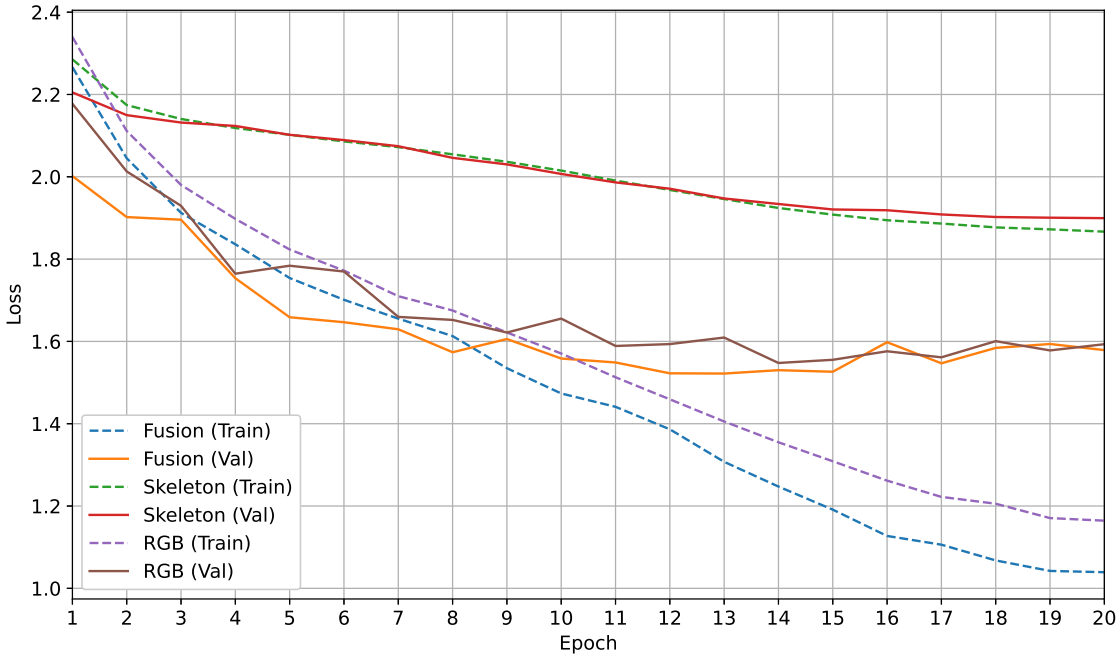

Figure 3. Combined loss plots from ablation study. Dashed lines indicate training data, while solid lines indicate validation data.

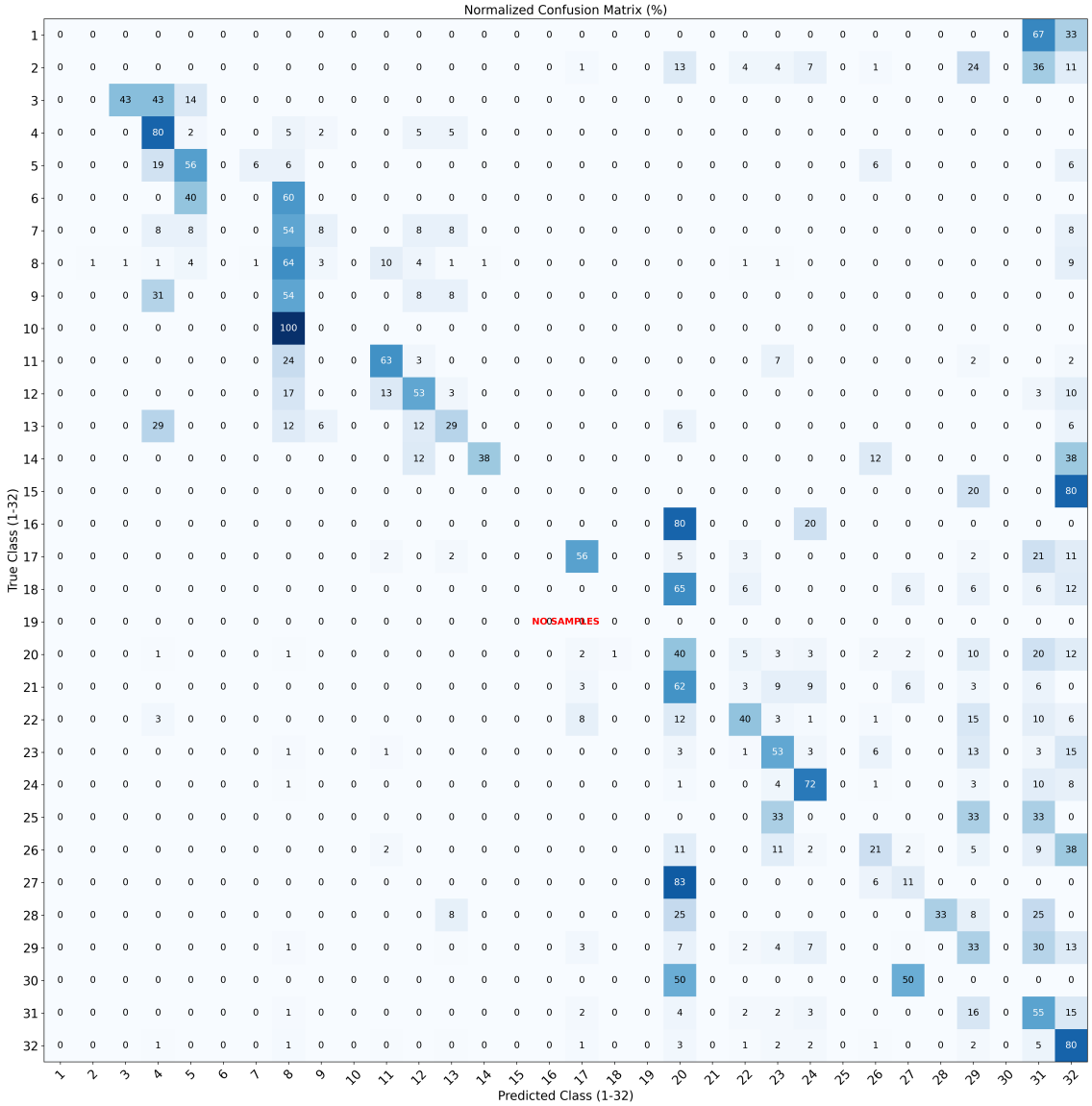

Figure 4. Normalized confusion matrix for the testing dataset using the fusion model. Class 19 contains no samples in the testing or validation set, hence the "no samples" text.

5 DISCUSSION

The experimental results (Figures 2 and 3) show clear differences in performance across the different methods. The fusion model consistently had the lowest training loss but also the highest accuracy. However, its marginal improvement over the RGB-only baseline suggests there is still room for better cross-modal interaction strategies. Both the RGB and fusion models indicate signs of overfitting, especially towards the end of training, as indicated by the increasing gap between validation and training loss values. In Figure 4, showing the confusion matrix, the diagonal values on majority classes show strong performance. However, some classes are notably missclassified, suggesting confusion with classes that share similar features.

The skeleton model's lower accuracy compared to benchmarks could stem from multiple factors. While OpenPose extraction errors, such as misdetected joints likely introduced noise, implementation choices such as graph convolutional layer design, insufficient temporal modelling or suboptimal hyperparameter tuning may have also limit performance. The skeleton model's bias toward majority classes suggests that it struggles with underrepresented actions, possibly due to limited training examples.

The fusion model's small improvement indicates that the current method does not fully integrate the strengths of both modalities. A likely reason is the simplistic fusion strategy, which may fail to align RGB and skeleton features effectively. Class imbalance further exacerbated these issues, particularly in the skeleton data, where minority classes lacked sufficient training examples.

Future work should explore architectural refinements such as adaptive graph convolutions, hyperparameter optimization and improved skeleton extraction methods. Testing alternative fusion strategies that may better align RGB and skeleton features over time. Addressing these factors could close the gap with state-of-the-art methods.

REFERENCES

[1] Xin Liu, Henglin Shi, Haoyu Chen, Zitong Yu, Xiaobai Li, and Guoying 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 (CVPR)*, pages 10631–10642, June 2021.

[2] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition, 2018.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. IEEE, 2016.

[4] Faisal Mehmood, Xin Guo, Enqing Chen, Muhammad Azeem Akbar, Arif Ali Khan, and Sami Ullah. Extended multi-stream temporal-attention module for skeleton-based human action recognition (har). *Computers in Human Behavior*, 163:108482, 2025.

[5] Hazdzz. Implementation of spatio-temporal graph convolutional networks. `https://github.com/hazdzz/stgcn/tree/main`, 2024. Accessed: [29.3.2025].