Recognizing Group Activities in Playground Scenes Using Multi-Person Graph Convolutional Networks (MP-GCN)

David Gómez
Tecnológico de Monterrey
Guadalajara, Mexico
a01642824@tec.mx

Angela Aguilar
Tecnológico de Monterrey
Guadalajara, Mexico
a01637703@tec.mx

Jorge Reyes
Tecnológico de Monterrey
Guadalajara, Mexico
a00573981@tec.mx

***Abstract*—Group Activity Recognition (GAR) in outdoor playground environments presents unique challenges due to occlusions, variable human poses, object-centric behaviors, and multi-person interactions. This paper presents a complete pipeline for recognizing group activities using 2D keypoints and a Multi-Person Graph Convolutional Network (MP-GCN). The system automatically performs frame extraction, pose estimation, multi-person tracking, CVAT-based annotations, and tensor construction for training. We classify scenes into three categories: Transit, Play_Object_Normal, and Play_Object_Risk. The experimental evaluation demonstrates stable learning tendencies in training accuracy and high variance in validation performance due to dataset imbalance and fine-grained behavior differences. Despite these constraints, MP-GCN shows strong potential for situational understanding and safety analysis in public play environments.**

***Index Terms*—Group Activity Recognition, MP-GCN, Skeleton-based Recognition, Video Understanding, Graph Neural Networks.**

I. Introduction

Group Activity Recognition (GAR) seeks to infer collective behavior from multi-person video sequences. Compared to single-person action recognition, GAR requires modeling interactions among multiple individuals and their surrounding objects. Recent surveys [3]–[5] highlight how skeleton-based representations and Graph Neural Networks (GNNs) have become preferred due to robustness, privacy considerations, and their ability to encode human motion dynamics.

Playground environments introduce unique challenges: irregular human motion, object manipulation, and frequent occlusions. RGB-based methods are heavily affected by lighting, background clutter, and appearance changes. Skeleton-based methods mitigate these issues by modeling pose geometry, following early multi-person interaction frameworks such as Choi et al. [2].

Graph Convolutional Networks (GCNs) have become foundational in skeleton-based recognition, especially after the introduction of Spatial Temporal GCNs (ST-GCN) by Yan et al. [6]. Subsequent advancements such as Adaptive GCNs [7] and explainable GCN baselines [8] have improved robustness. Multi-Person GCNs further extend these architectures by incorporating human–human and human–object interactions, as demonstrated by Li et al. [1].

In this work, we adapt the MP-GCN architecture for playground safety classification and develop a full video-to-graph dataset pipeline.

II. Related Work

Skeleton-based Human Action Recognition (HAR) has expanded rapidly. Surveys such as Feng and Meunier [3], Li et al. [4], and Liu et al. [5] provide comprehensive overviews of GNN-based action recognition, emphasizing their advantages over CNN and RNN models.

The breakthrough ST-GCN architecture by Yan et al. [6] formalized the use of graph convolutions over skeleton sequences. Later improvements include the Two-Stream Adaptive GCN [7], which introduced learnable adjacency matrices, and the Stronger, Faster, More Explainable GCN baseline by Song et al. [8]. Hybrid temporal models such as Attention-Enhanced GCN-LSTM [9] further demonstrated benefits of temporal attention.

Earlier models such as hierarchical RNNs [10] and compact 3D sequence representations [11] contributed important insights on motion structure. Datasets like NTU RGB+D [12] set modern benchmarks for skeleton-based recognition.

Our work differs from prior HAR research by targeting GAR in playground environments, incorporating both human–human and human–object interactions via a panoramic multi-person graph as in Li et al. [1].

III. Methodology

A. Video Processing and Frame Extraction

Videos are sampled at 15 FPS. Frames are extracted using ffmpeg and stored as JPEG sequences.

B. Skeleton Detection and Tracking

YOLO-Pose extracts 17-joint COCO skeletons. DeepSort provides consistent identity tracking, similar to pipelines described in [3].

C. Object Detection and Context Modeling

Playground objects such as slides and swings are annotated manually in CVAT. These centroids are used as static contextual nodes in the graph. We currently treat these objects as static and do not apply temporal smoothing.

D. Annotation Using CVAT

Scene-level labels include Transit, Play_Object_Normal, and Play_Object_Risk.

E. Tensor Construction

Each clip is converted into:

$$X \in \mathbb{R}^{2\times 30\times 21\times 6}$$

where human joints and object centroids form the graph nodes.

IV. Dataset Construction Details

Pose filtering removes low-confidence joints and anatomically implausible limb proportions. Missing joints are interpolated temporally [4]. Object annotations combine automated proposals with manual refinement. Centroids are tracked across frames to capture consistent interaction cues. The resulting dataset is a structured representation combining pose, context, and temporal information.

V. Graph Formulation and Adjacency Design

The adjacency matrix includes:

- Intra-person edges (skeletal topology [6]),
- Inter-person edges (proximity-based),
- Person–object edges (human–context interactions).

Inspired by adaptive adjacency works [7], we integrate learnable adjacency refinement to dynamically weight important edges.

VI. MP-GCN Architecture

MP-GCN extends ST-GCN-like spatial–temporal blocks with multi-person and human–object connectivity. Temporal convolutions capture motion patterns, supported by findings in GCN-LSTM hybrids [9].

VII. Training Setup and Hyperparameters

Models are trained with Adam (learning rate 0.001) for 30 epochs. Dropout and weight decay reduce overfitting. [5].

VIII. Experiments

A. Dataset Characteristics

- Transit: 74 samples,
- Play_Object_Normal: 25 samples,
- Play_Object_Risk: 21 samples.

B. Augmentation

We apply temporal jittering, Gaussian noise, person dropout, and random scaling.

C. Training Performance

Fig. 1 shows accuracy curves. Training accuracy increases steadily to $\sim$ 0.62, while validation accuracy fluctuates strongly—consistent with observations in MP-GCN and GNN literature on small datasets [3], [5].

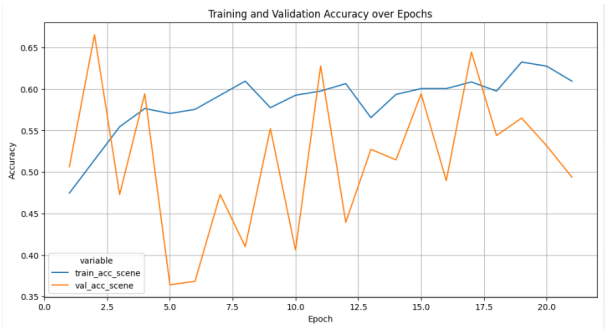

Fig. 1. Training and Validation Accuracy over Epochs.

Beyond overall accuracy trends, several finer observations can be made. The smooth, monotonic ascent of training accuracy indicates that MP-GCN consistently extracts discriminative patterns from spatial–temporal graph structures, echoing behavior documented in ST-GCN and AGCN architectures [6], [7].

Validation instability reflects the strong class imbalance and subtle pose variations between normal and risky interactions. Similar oscillations are reported in skeleton-based recognition surveys [3], [4]. These fluctuations may stem from minor pose-tracking noise, occlusions, or similarities in climbing-related motions across classes.

Moreover, despite the small dataset, the model does not overfit. Dropout, GraphBatchNorm, and augmentation strategies play a key role in maintaining generalization. However, the validation noise suggests that discriminative boundaries remain poorly separated, motivating future integration of temporal attention [9] or richer context features [8].

D. Confusion Matrix Analysis

Fig. 2 shows the confusion matrix. Transit achieves the highest recognition rate, which is expected given its distinctive locomotion pattern and strong temporal cues. This mirrors patterns observed in benchmarks like NTU RGB+D [12], where walking-related classes show consistently higher recall.

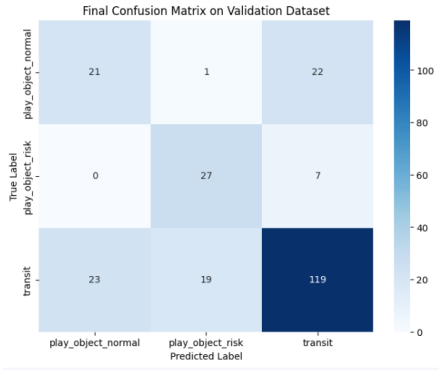

Fig. 2. Confusion Matrix on Validation Set.

Normal and risk-related object classes exhibit higher confusion. Both involve interaction with structures (slides, ramps, swings), producing overlapping motion signatures. This challenge is widely recognized in fine-grained skeleton-based HAR [11].

Play_Object_Risk is the most difficult category, frequently misclassified as Normal. Subtle risk cues—lean angle, support loss, rapid transitions—may not be fully captured by 2D joint geometry alone. Contextual enhancement (object motion, depth information) has been shown to significantly improve discrimination in related tasks [8].

These findings highlight the need for additional sensory or semantic cues to fully separate fine-grained behaviors.

E. Per-Class Precision, Recall, and F1-Score

To further analyze class-specific behavior, Fig. 3 reports the precision, recall, and F1-score for each activity category.

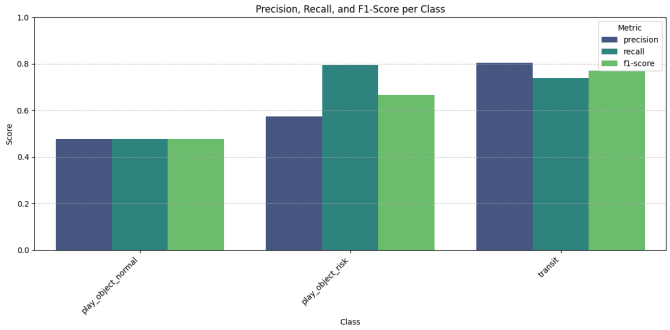

Fig. 3. Precision, Recall, and F1-Score per Class.

Transit achieves the best overall performance, with the highest precision (0.80) and F1-score (0.77), and competitive recall (0.74). This agrees with findings in skeleton-based datasets such as NTU RGB+D, where locomotion-related actions often exhibit strong separability. Both object-related classes show substantially lower precision than transit. Play Object Normal achieves balanced yet modest scores (around 0.48), reflecting ambiguity in body posture and interaction context. Play Object Risk displays the highest recall (0.80) but only moderate precision (0.57), suggesting that while the model detects most risky cases, it also produces many false positives. This mirrors limitations documented in fine-grained HAR research, where subtle pose and balance cues are difficult to capture from 2D skeletons alone.

The F1-score trends confirm that recognizing risk-related behavior remains challenging and likely requires richer contextual and temporal information. Recent works emphasize the need for adaptive graph representations or temporal attention mechanisms, which may help improve discrimination between normal and risky object interaction.

IX. Limitations and Future Work

Limitations include dataset imbalance, reliance on 2D pose, and static object modeling. Surveys [3], [5] point to benefits of 3D pose, multi-view inputs, and attention-based temporal modeling [9].

Future directions include dynamic object nodes, 3D skeletal cues, semi-supervised learning, and multiscale temporal attention.

Conclusions

This work presents a full GAR pipeline using MP-GCN for playground environments. By integrating skeletons, object context, and multi-person interactions, our method aligns with trends highlighted in recent GNN-based HAR surveys [3]–[5]. Results demonstrate feasibility for safety classification, with opportunities for improvement through richer temporal and contextual modeling.

References

[1] Z. Li, X. Chang, Y. Li, and J. Su, "Skeleton-Based Group Activity Recognition via Spatial-Temporal Panoramic Graph," in *LNCS*, 2024.
[2] W. Choi, K. Shahid, and S. Savarese, "What are they doing?: Collective activity classification using spatio-temporal relationships among people," in *ICCV Workshops*, 2009.
[3] M. Feng and J. Meunier, "Skeleton graph-neural-network-based human action recognition: A survey," *Sensors*, 2022.
[4] C. Li, Y. Hou, P. Wang, and W. Li, "Skeleton graph neural network for human action recognition: A survey," *IJCV*, 2021.
[5] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and R. Ji, "Deep learning for skeleton-based human action recognition: A survey," *IEEE TPAMI*, 2020.
[6] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *AAAI*, 2018.
[7] S. Yan, Y. Xiong, and D. Lin, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *CVPR*, 2019.
[8] Y.-F. Song, Z. Zhang, C. Shan, and L. Wang, "Stronger, faster and more explainable: A GCN baseline for skeleton-based action recognition," in *ACM MM*, 2019.
[9] C. Si, Y. Jing, W. Wang, L. Wang, and T. Tan, "Attention enhanced graph convolutional LSTM network for skeleton-based action recognition," in *CVPR*, 2019.
[10] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton-based action recognition," in *CVPR*, 2015.
[11] Q. Ke, S. An, M. Bennamoun, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3D action recognition," in *CVPR*, 2017.
[12] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3D human activity analysis," in *CVPR*, 2016.