Interaction-Aware Multimodal Spatiotemporal Learning For Fine-Grained Volleyball Training Action Recognition

Jing Wang[1], Haojuan Wang[2*], and Jinglong Geng[3]

[1] *Shanxi Vocational University of Engineering Science and Technology, Physical Education College, Shanxi Jinzhong 030619, China*
[2] *Shanxi University of Chinese Medicine, Shanxi Jinzhong 030619, China*
[3] *Taiyuan Normal University, Shanxi Jinzhong 030619, China*
[*] *Corresponding author. E-mail: wang22007@outlook.com*

Received: May. 12, 2026; Accepted: Jul. 07, 2026

To address the challenges of significant fine-grained action differences, complex multi-agent interactions, and strong temporal dynamic dependencies in volleyball scenarios, this paper proposes an action recognition method based on multimodal spatiotemporal fusion. First, convolutional neural networks are used to extract spatial features from images, and human pose information is combined to enhance the representation of fine-grained structural differences. Second, a hierarchical temporal modeling framework is constructed, using individual-level LSTM and group-level LSTM to characterize the action evolution of individual players and multi-agent interaction relationships, respectively. Simultaneously, a People Pooling mechanism is introduced to adaptively fuse features from different subjects, highlighting key roles and modeling "ball-player-player" interactions. Finally, RGB and Skeleton multimodal information are fused to improve model robustness. Experimental results on the publicly available Volleyball dataset and a self-built dataset demonstrate that the proposed method outperforms several baseline methods in terms of accuracy, F1-score, and mAP, validating its effectiveness and generalization ability in fine-grained action recognition and complex scene modeling.

Keywords: Volleyball motion recognition; fine-grained motion modeling; multi-agent interaction; temporal modeling; multimodal fusion

© The Author('s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are cited.

http://dx.doi.org/10.6180/jase.202611_34.060

1. Introduction

With the rapid development of computer vision and deep learning technologies, videobased action recognition has become increasingly important for intelligent sports analysis, training assistance, and human-computer interaction [1]. Accurate volleyball action recognition enhances match analysis, tactical evaluation, and athlete training in complex scenarios [2, 3]. Compared to traditional single-player action recognition tasks, volleyball action recognition is significantly more complex. Volleyball involves multiple athletes whose actions are highly interactive, requiring coordinated behaviors such as attacking, passing, and blocking. In addition, fine-grained actions, including reception and setting, exhibit similar postures but differ in subtle joint movements and action intent. Moreover, action semantics rely on continuous temporal evolution rather than isolated frames. Therefore, effectively modeling fine-grained spatial features, temporal dynamics, and multi-agent interactions remains a fundamental challenge for achieving accurate volleyball action recognition [4, 5].

To address the aforementioned issues, numerous research efforts have been undertaken in recent years [6]. Early handcrafted feature-based methods showed limited performance, while CNN-based approaches improved spatial feature extraction but struggled to capture temporal

action dynamics due to reliance on single-frame information [7, 8].

To solve the temporal modeling problem, researchers have introduced recurrent neural networks (RNNs) and variants such as LSTM and GRU for video sequence modeling, while 3D convolution methods (e.g., I3D) capture spatiotemporal features. However, multi-subject interaction modeling remains limited [9, 10]. Skeleton-based action recognition methods have gradually emerged. Graph Convolutional Networks (GCNs) and ST-GCNs effectively model human joint topologies for fine-grained action recognition [11, 12]. However, they primarily capture individual structural information while overlooking multi-subject interactions and contextual scene relationships.

For the problem of group behavior recognition, some studies have proposed hierarchical modeling methods (such as Hierarchical LSTM) to improve recognition performance by modeling individual and group behaviors separately [13]. However, these methods often have limited capability in capturing fine-grained spatial differences and complex multi-agent interactions. In addition, most existing approaches rely on a single modality (e.g., RGB or Skeleton), limiting the exploitation of complementary multimodal information. Consequently, current methods still face challenges in distinguishing similar actions, modeling temporal dynamics, representing multi-subject interactions, and effectively fusing RGB and pose features for robust action recognition [14, 15].

To address the aforementioned issues, this paper proposes a multimodal spatiotemporal fusion method for volleyball motion recognition. The method integrates fine-grained modeling, multi-agent interaction, and temporal dynamics within a unified framework. Spatial features extracted by a convolutional neural network are combined with human pose information to enhance structural representation. A hierarchical LSTM models individual motion evolution and group interactions, while a People Pooling mechanism adaptively aggregates player features to emphasize key roles and ball-player interactions. Finally, RGB and Skeleton modalities are fused to obtain robust spatiotemporal representations, improving recognition accuracy and generalization in volleyball scenarios.

2. Materials and methods

2.1. Preliminary

2.1.1. *Fine-grained motion difference representation based on human posture*

Human pose estimation represents the body as skeletal key joints, improving fine-grained sports action recognition. Focusing on local joint structures rather than global appearance, pose-based representations suppress background interference and provide discriminative motion cues, enabling accurate differentiation of visually similar volleyball actions. As shown in Fig. 1, joint-position variations distinguish fine-grained volleyball actions despite identical skeletal topology:

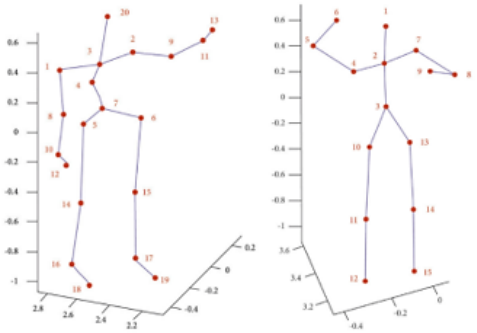

Fig. 1. Skeletal representation of fine-grained action states

$$\mathcal{G}^{(L)} = \left(\mathcal{V}^{(L)}, \mathscr{E}\right), \qquad \mathcal{G}^{(R)} = \left(\mathcal{V}^{(R)}, \mathscr{E}\right) \tag{1}$$

Where, $\mathcal{V}$ and $\mathscr{E}$ denote joints and skeletal connections, respectively; identical topology with different coordinates represents distinct motion states given in **??**. In the skeletal representation, each node corresponds to a human body joint, while edges describe anatomical connections, forming a structured human body topology. This graph-based representation preserves spatial joint locations and inter-joint relationships, providing a robust foundation for fine-grained motion analysis. To improve robustness, confidence scores assess the reliability of detected keypoints, reducing the influence of inaccurate localizations. Consequently, the skeletal representation remains stable under challenging conditions such as occlusion, motion blur, and complex player interactions, supporting reliable volleyball action recognition. Based on this, to characterize the fine-grained spatial differences between the left and right motions, joint-level difference modeling is introduced in Eq. (2) as:

$$\Delta\mathcal{P} = \sum_{i=1}^{N} w_i \cdot \left| v_i^{(L)} - v_i^{(R)} \right| \tag{2}$$

Where $v_i^{(L)}$ and $v_i^{(R)}$ denote left/right joint positions, and w_i is the joint importance weight emphasizing discriminative volleyball joints. All joint importance weights are initialized uniformly and updated during training through backpropagation together with the network parameters. This adaptive weighting mechanism allows the model to

progressively emphasize joints that contribute most significantly to the discrimination of visually similar volleyball actions.

The joint-level difference term $\Delta\mathcal{P}$ further enables the model to identify discriminative structural variations associated with specific volleyball actions. For instance, reception, setting, and attacking actions exhibit distinct movement patterns in the upper-limb joints, and modeling these fine-grained joint differences improves the recognition of visually similar actions.

2.2. Multi-Agent Interaction and Temporal Dynamics in Volleyball Scenes

Volleyball motion recognition requires modeling multi-subject and ball interactions beyond individual posture to accurately capture action semantics in complex scenarios.

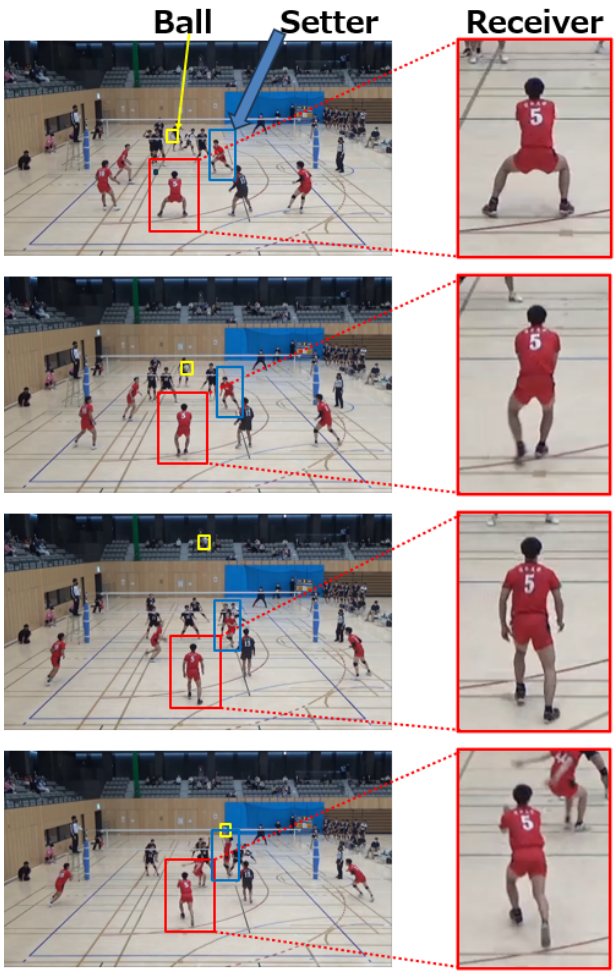

Fig. 2. Multi-agent interaction and temporal evolution in a volleyball scenario

2.3. Volleyball Motion Recognition Framework

2.3.1. *Overall Model Framework*

Based on fine-grained pose and multi-subject interaction analysis, Fig. 3 presents a multimodal spatiotemporal fusion framework integrating CNN spatial extraction, LSTM temporal modeling, and feature fusion for volleyball recognition.

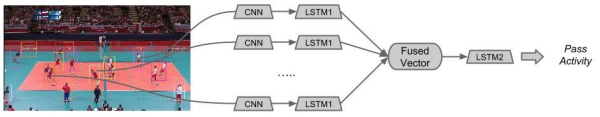

Fig. 3. Multimodal spatiotemporal fusion framework

Given an input video sequence in Eq. (3) as:

$$V = I_{t=1}^{T} \tag{3}$$

Where I_t denotes the input image at frame t. Frame-wise features are first extracted and then temporally modeled to generate a comprehensive action representation. Specifically, the proposed framework integrates CNN-based spatial feature extraction, hierarchical LSTM-based temporal modeling at both the individual and group levels, and RGB–skeleton multimodal fusion for robust fine-grained volleyball action classification. The CNN extracts spatial pose features, whereas the LSTM captures their temporal evolution for volleyball action recognition (Fig. 3 and Eqs. (4) to (6)). For the input feature sequence $\{\mathbf{F}_t\}_{t=1}^{T}$, the temporal representation is defined as:

$$\boldsymbol{H} = \mathrm{LSTM}\left(\mathbf{F}_t\right) \tag{4}$$

This process models temporal dynamics $\Delta\mathcal{T}$, capturing action evolution and distinguishing actions with similar spatial structures. The temporal dynamic dependency $\Delta\mathcal{T}$ captures the sequential evolution of volleyball actions across preparation, movement, and execution phases. Modeling these transitions enables learning of action-specific temporal patterns, improving discrimination between actions with similar spatial postures but different temporal characteristics. The feature fusion pathway connects CNN-based spatial encoding and hierarchical sequence learning. Spatial features are transformed through individual-level LSTMs and aggregated for group-level modeling, enabling

unified learning of spatial structures, temporal dynamics, and interaction semantics for accurate volleyball action recognition. As shown in Fig. 3, volleyball action recognition integrates player, ball, and setter interactions through unified feature fusion. Let the feature representation corresponding to different branches be $\boldsymbol{H}^{(k)}{}_{k=1}^{K}$, then the fusion process can be represented as:

$$\boldsymbol{H}_{\text{fused}} = \sum k = 1^{K} \alpha_k \boldsymbol{H}^{(k)} \tag{5}$$

Where α_k denotes adaptive branch weights, emphasizing key players, while ball position provides essential contextual information. In the proposed multimodal spatiotemporal fusion framework, RGB features capture appearance and scene context, while skeleton features represent fine-grained joint structures. During temporal modeling, both modalities jointly characterize motion dynamics and player interactions across consecutive frames. Their complementary representations are subsequently fused for final action recognition, enabling more robust and accurate volleyball action classification. Effective multimodal integration is achieved by synchronizing RGB frames and corresponding skeleton sequences at the frame level prior to feature extraction. The extracted features are then aligned within a unified feature space and jointly learned during temporal modeling and fusion, enabling the complementary exploitation of appearance, pose, and interaction information for action recognition. The proposed framework integrates fine-grained modeling, multi-agent interaction, and temporal dynamics within a unified spatiotemporal representation. Fine-grained features capture subtle action differences, temporal modeling characterizes motion evolution, and multi-agent fusion encodes collaborative relationships, jointly supporting accurate action recognition. By integrating spatial, temporal, and multiagent modeling, the proposed method accurately recognizes complex volleyball actions.

2.3.2. *Hierarchical Temporal Modeling for Person-Group Dynamics*

As shown in Fig. 4, a hierarchical temporal model captures individual motion evolution and multi-agent collaboration. Hierarchical LSTM was selected because it effectively models person-level and group-level temporal dependencies in volleyball interactions. Compared with Transformer-based architectures, it requires fewer training samples and lower computational cost while effectively capturing sequential motion continuity and fine-grained interaction dynamics, making it suitable for multi-agent action recognition.

Adaptive weights aggregate individual dynamics into group representation, emphasizing action-relevant subjects (e.g., the catching player).

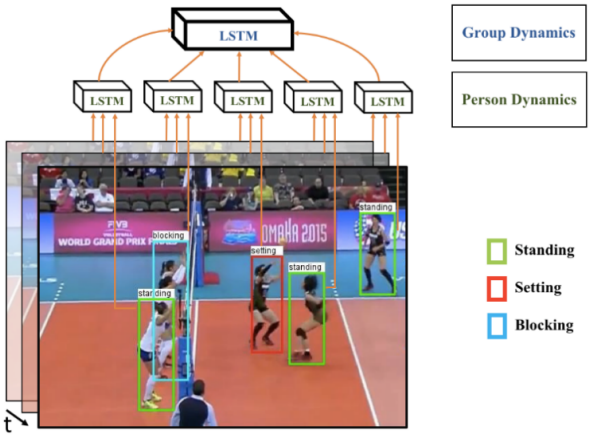

Fig. 4. Fine-grained and interaction-aware temporal modeling

$$g_t = \sum i = 1^{N} \alpha_{i,t}, \boldsymbol{h}_{i,t}, \quad \alpha_{i,t} = \frac{\exp\left(\psi\left(\boldsymbol{h}_{i,t}\right)\right)}{\sum j = 1^{N} \exp\left(\psi\left(\boldsymbol{h}_{j,t}\right)\right)} \tag{6}$$

Where $\psi(\cdot)$ is a learnable scoring function and $\alpha_{i,t}$ denotes normalized weights emphasizing key interactions. The subject scores reflect the relative importance of each player to the current action and are used to adaptively weight individual features during aggregation. This mechanism enhances interaction-based feature extraction by emphasizing key agents and capturing collaborative relationships among multiple players within the group activity. Consequently, players assigned higher values of $\alpha_{i,t}$ contribute more strongly to the aggregated group representation g_t, allowing actioncritical participants to dominate the interaction modeling process and enhancing recognition of coordinated volleyball activities.

The people pooling process aggregates temporal representations from all detected players into a unified group descriptor. Unlike simple averaging, adaptive weighted fusion assigns greater importance to action-relevant roles, such as receivers and setters, enabling more effective representation of collective behavior in dense volleyball scenes. During multi-agent feature aggregation, player features are organized according to their detected instances within each frame prior to processing by the People Pooling module, ensuring consistency in sequence encoding. The aggregation mechanism emphasizes the relevance of individual player features rather than their positional ordering, allowing the model to capture interaction patterns while reducing sensitivity to variations in player arrangement across different volleyball scenarios. The hierarchical temporal modeling framework first employs individual-level LSTMs to learn

the temporal evolution of each player's actions. The resulting person-level representations are then aggregated and provided to the group-level LSTM, which captures collective dynamics and coordination patterns among multiple players for group activity recognition. This hierarchical representation facilitates the transition from individual behavior modeling to group semantic understanding. By integrating player-specific temporal patterns and inter-player interactions, the framework captures collective activity semantics that cannot be inferred from isolated individual actions alone.

2.3.3. *Spatial Feature Encoding and Spatiotemporal Aggregation Modeling*

Fig. 5 shows CNN-based multi-layer feature extraction for multi-scale spatial representations in volleyball motion recognition.

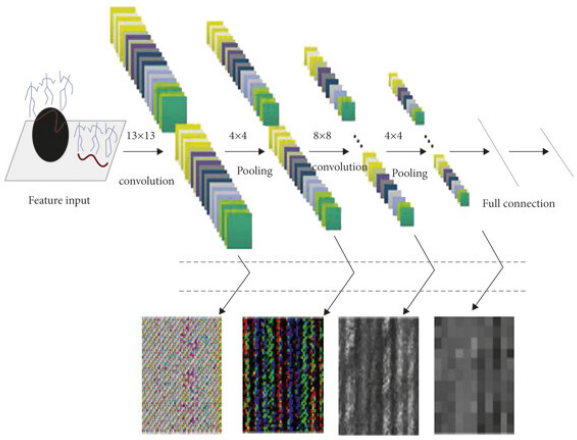

Fig. 5. CNN-based spatial feature extraction pipeline

The hierarchical multi-scale feature extraction captures local joint details and global action semantics, enabling accurate fine-grained volleyball action recognition.

3. Result & discussion

3.1. Experimental Setup

To verify the effectiveness and generalization ability of the proposed method, experiments were conducted on a public dataset and a self-built dataset. The CNN backbone uses 512-dimensional feature embeddings, while Person LSTM and Group LSTM use 256 hidden units. The temporal sequence length is fixed to 16 frames for both training and inference to preserve motion continuity. Training uses a batch size of 32 and a learning rate of 0.001 . Dataset samples are divided into training, validation, and testing sets with a 70:10:20 ratio: Volleyball Dataset (3381/483/966) and self-built dataset (4200/600/1200). Subject-independent evaluation ensures generalization, while RGB-skeleton fusion and hierarchical temporal modeling improve feature representation and enable stable convergence.

3.1.1. *Datasets*

- (1) Volleyball Dataset

The Volleyball Dataset contains 55 match videos and approximately 4830 labeled frames with actions such as pass, set, spike, and block. It provides player positions and labels, supporting multi-agent interaction and temporal analysis evaluation.

- (2) Volleyball Training Dataset

The proposed volleyball training dataset includes four motion categories from real videos with complex backgrounds (Fig. 6). Data were collected under diverse real-world conditions using multi-view cameras, incorporating varying viewing angles, player scales, occlusions, lighting, backgrounds, and recordings from training and competitive matches to evaluate model generalization. Each frame is annotated with motion categories and key player regions, validating temporal modeling under complex scenes.

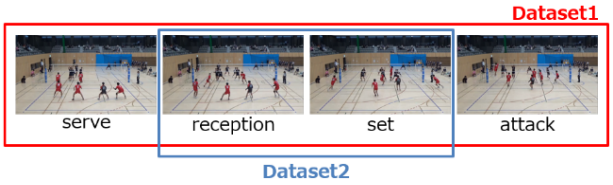

Fig. 6. Self-built dataset

3.2. Comparison with State-of-the-Art Methods

To evaluate performance, the proposed method was compared with representative baselines, including single-frame, temporal, skeleton, and group behavior models, on Volleyball Dataset (Vol) and Proposed Dataset (Prop).

As shown in Table 1, the proposed method achieves superior performance across both datasets, improving Accuracy/F1 by +3.3%/+2.9% (Vol) and +3.7%/+3.9% (Prop), while attaining the highest mAP (92.3%, 89.7%). Cross-dataset evaluation on the Prop dataset shows minimal performance degradation ($1.4\% - 1.8\%$), demonstrating strong robustness and generalization under complex variations. The proposed framework maintains practical computational efficiency through hierarchical temporal modeling and linear-complexity People Pooling, reducing computational overhead and improving inference-time performance for efficient near-real-time volleyball action analysis.

Table 1. Comparison of Performance on Volleyball and Proposed Datasets

Method	Volleyball Dataset				Proposed Dataset			
	Acc	F1	Top-5	mAP	Acc	F1	Top-5	mAP
CNN	82.3	80.5	91.2	78.6	78.6	76.9	88.4	74.2
CNN+LSTM	86.7	85.2	93.5	83.7	83.2	81.5	90.6	79.8
Two-Stream	88.4	87.6	94.8	86.5	85.1	83.7	92.1	81.9
I3D	89.0	88.1	95.2	87.4	86.0	84.6	92.8	82.7
Pose-GCN	89.1	88.3	95.4	87.9	86.8	85.2	93.1	83.4
ST-GCN	90.0	89.2	96.0	88.6	87.9	86.4	93.9	84.5
Hierarchical LSTM	90.5	89.7	96.3	89.1	88.7	87.1	94.2	85.3
Ours	**93.8**	**92.6**	**97.8**	**92.3**	**92.4**	**91.0**	**96.5**	**89.7**

Statistical significance was verified using multiple runs and a paired Student's t-test. Table 2 shows significant improvements ($p < 0.05$) in Accuracy, F1-score, Top-5 Accuracy, and mAP.

3.3. Ablation Study

Ablation experiments progressively evaluated proposed modules on the public Vol and self-built Prop datasets, validating model effectiveness, stability, and generalization.

As shown in Table 3, progressive module integration improves performance: LSTM captures temporal dynamics, Person LSTM models individual behavior, Group LSTM learns interactions, and People Pooling enhances feature aggregation. These results demonstrate the complementary contributions of the proposed components. Hierarchical Person and Group LSTMs capture individual and group dynamics, People Pooling emphasizes action-relevant players through adaptive aggregation, and

RGB+Skeleton fusion integrates appearance and pose information for improved finegrained volleyball action recognition. Finally, the model achieves state-of-the-art performance on both datasets, demonstrating complementarity between the modules. A numerical modality contribution analysis showed that RGB + Skeleton fusion consistently outperformed individual modalities, demonstrating complementary feature interactions for volleyball action recognition Table 4.

3.4. Visualization Results

As shown in Fig. 7, the proposed method detects players and robustly captures finegrained joint differences (ΔP) under occlusion and viewpoint variations. The framework also addresses severe occlusion and overlapping athletes through the integration of pose-based representations and hierarchical temporal modeling. When body regions are partially obscured, temporal dependencies across consecutive frames and multi-agent interaction cues help maintain action consistency, while multimodal feature fusion

Fig. 7. Visualization of pose estimation and multi-agent interaction modeling

enhances resilience to viewpoint variations. The model captures multiplayer positional relationships through Group LSTM and People Pooling to model interaction relationships $\mathscr{L}_{\text{inter}}$. Temporal modeling $\Delta\mathcal{T}$ maintains consistent action understanding across frames, supporting accurate action classification.

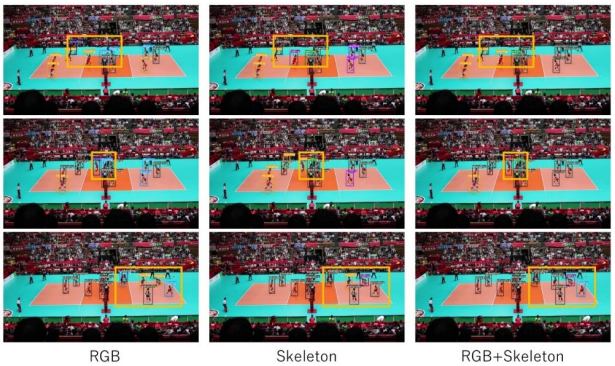

Fig. 8. RGB, Skeleton, and fused feature representations comparison

As shown in Fig. 8, fusion outperforms RGB by combining contextual and joint features, improving recognition accuracy. The effectiveness of multimodal fusion arises

Table 2. Statistical significance comparison with the best baseline

Metric	Proposed Mean ± SD	Best Baseline Mean ± SD	p value	Significant
Accuracy	93.80 ± 0.28	90.50 ± 0.34	<0.001	Yes
F1-score	92.60 ± 0.31	89.70 ± 0.37	<0.001	Yes
Top-5 Accuracy	97.80 ± 0.15	96.30 ± 0.22	0.002	Yes
mAP	92.30 ± 0.27	89.10 ± 0.36	<0.001	Yes

Table 3. Ablation results on Vol and Prop datasets

Model Variant	Acc (Vol)	F1 (Vol)	Acc (Prop)	F1 (Prop)
Baseline (CNN)	82.3	80.5	78.6	76.9
+ LSTM	86.7	85.2	83.2	81.5
+ Person LSTM	89.2	88.0	85.9	84.3
+ Group LSTM	91.0	89.8	87.6	86.2
+ People Pooling	92.4	91.1	89.3	88.0
Full Model	93.8	92.6	92.4	91.0

Table 4. Numerical modality contribution analysis

Input Modality	Accuracy	F1-score	mAP
RGB Only	90.8	89.5	88.2
Skeleton Only	88.9	87.8	86.4
RGB + Skeleton Fusion	93.8	92.6	92.3

Table 5. Quantitative localization and skeletal reliability evaluation

Metric	Value
PCK@0.05	94.80%
PCK@0.10	97.10%
OKS	0.931
Mean Joint Error (pixels)	4.12
Occlusion Localization Accuracy	91.60%
Skeleton Reliability Score	95.40%

from the complementary information provided by RGB and skeleton representations. RGB features capture appearance and contextual details, while skeleton features represent body structure and motion patterns. Their integration reduces feature redundancy and enhances discriminative capability. This fusion strategy improves representation quality, enabling more reliable recognition of ambiguous volleyball actions under complex conditions by combining visual context with motion and posture cues.

Quantitative localization analysis using PCK, OKS, Mean Joint Error, and Skeleton Reliability Score validates spatial consistency and robust pose recognition Table 5.

4. Conclusion

This paper addresses the challenges of fine-grained movement differences, complex multi-agent interactions, and strong temporal dynamic dependencies in volleyball scenarios, proposing a multimodal spatiotemporal fusion-based action recognition method that improves action recognition in complex scenarios. Specifically, this paper focuses on fine-grained modeling, combining RGB visual features with human posture information to represent differences in joint structures, thereby enhancing the ability to distinguish similar action categories. Building upon this, a hierarchical temporal modeling framework is constructed, using individual-level LSTM and group-level LSTM to model the evolution of individual players' movements and collaborative relationships between multiple agents, elevating the model from "individual behavior" to "group semantics." Furthermore, this paper introduces a People Pooling mechanism to adaptively weight and fuse features from agents, strengthening the expressive power of key roles and effectively modeling ball-player interactions. Experimental results show that the proposed method outperforms several baseline methods on both the Volleyball dataset and the self-built dataset, demonstrating improvements in accuracy, F1-score, Top-5, and mAP. Ablation experiments validate the effectiveness of each module, while visualization analysis demonstrates the model's robustness and stability in complex scenes. Although the

proposed method achieves good performance in volleyball motion recognition, it still has limitations. For example, the model is somewhat dependent on target detection accuracy and may still be affected in extreme occlusion or dense scenes; in addition, the current method is mainly based on two-dimensional visual information, and the modeling of three-dimensional spatial relationships needs further improvement. Future work will consider introducing a Transformer structure to enhance long-term dependency modeling capabilities and combining 3D pose information with multi-view data to further improve the model's performance and generalization ability in complex real-world scenes.

5. Acknowledgements

The authors would like to show sincere thanks to those techniques who have contributed to this research.

6. Declaration:

7. Ethics approval

Not applicable

8. Consent to participate

Not applicable

9. Consent for publication

Not applicable

10. Competing interests

The authors declare no competing interests.

11. Funding

There is no specific funding to support this research.

12. Author contributions

Jing Wang,Haojuan Wang ,Jinglong Geng conceptualized the study, designed its framework, and drafted and finalized the manuscript.

13. Data availability

The experimental data used to support the findings of this study are available from the corresponding author upon request.

References

[1] I. Zahra, Y. Wu, H. F. Alhasson, S. S. Alharbi, H. Aljuaid, A. Jalal, and H. Liu, (2025) *"Dynamic graph neural networks for UAV-based group activity recognition in structured team sports"* **Frontiers in Neurorobotics *19***: 1631998. DOI: 10.3389/fnbot.2025.1631998.

[2] A. Alqarafi and B. Almogadwy, (2025) *"STRIKE-net: An explainable dynamic spatiotemporal graph-transformer network for fine-grained soccer action recognition"* **Applied Soft Computing**: 114224. DOI: 10.1016/j.asoc.2025.114224.

[3] K. Seweryn, A. Wróblewska, and S. Łukasik, (2026) *"Survey of action recognition, spotting, and spatio-temporal localization in soccer—current trends and research perspectives"* **ACM Transactions on Intelligent Systems and Technology *17***(2): 1–37. DOI: 10.1145/3776541.

[4] C. Zhang, G. Shan, J. Lim, and B. H. Roh, (2024) *"Dynamic reinforcement learning for optimal go AI training: adaptive adjustment and optimization"* **IEEE Transactions on Consumer Electronics *71***(1): 292–302. DOI: 10.1109/TCE.2024.3487141.

[5] M. A. Hossen and P. E. Abas, (2025) *"Machine learning for human activity recognition: State-of-the-art techniques and emerging trends"* **Journal of Imaging *11***(3): 91. DOI: 10.3390/jimaging11030091.

[6] N. V. S. R. Chappa and K. Luu. "LiGAR: LiDAR-guided hierarchical transformer for multi-modal group activity recognition". In: *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2025, 3035–3044. DOI: 10.1109/WACV61041.2025.00300.

[7] D. Karki, M. Ramazanova, A. Cioppa, S. Giancola, and B. Ghanem. "Pixels or positions? benchmarking modalities in group activity recognition". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2026, 9998–10008. DOI: 10.48550/arXiv.2511.12606.

[8] L. Wang, P. Koniusz, and Y. Gao, (2025) *"Video Understanding by Design: How Datasets Shape Architectures and Insights"* **arXiv preprint arXiv:2509.09151**: DOI: 10.48550/arXiv.2509.09151.

[9] R. Zhang, Y. Wang, X. Hu, C. Mai, W. Liu, D. Xu, et al. "Beyond the Individual: Introducing Group Intention Forecasting with SHOT Dataset". In: *Proceedings of the 33rd ACM International Conference on Multimedia*. 2025, 13002–13008. DOI: 10.1145/3746027.3758248.

[10] Y. Liu, F. Liu, L. Jiao, Q. Bao, L. Li, Y. Guo, and P. Chen, (2024) *"A knowledge-based hierarchical causal inference network for video action recognition"* **IEEE Transactions on Multimedia *26***: 9135–9149. DOI: 10.1109/TMM.2024.3386339.

[11] Z. Chen, W. Sun, Y. Tian, J. Jia, Z. Zhang, J. Wang, et al. "Gaia: Rethinking action quality assessment for AI-generated videos". In: *Advances in Neural Information Processing Systems*. ***37***. 2024, 40111–40144. DOI: 10.52202/079017-1267.

[12] X. Wang, X. Lan, and W. Zhu. *Video Grounding and Its Generalization: From ID and Task-specific Models to OOD and Large Foundation Models*. Springer International Publishing AG, 2025. DOI: 10.1007/978-3-031-94837-4.

[13] Y. Wen, M. Liu, S. Wu, and B. Ding. "Chase: Learning convex hull adaptive shift for skeleton-based multi-entity action recognition". In: *Advances in Neural Information Processing Systems*. ***37***. 2024, 9388–9420. DOI: 10.52202/079017-0298.

[14] B. Amirgaliyev, M. Mussabek, T. Rakhimzhanova, and A. Zhumadillayeva, (2025) *"A review of machine learning and deep learning methods for person detection, tracking and identification, and face recognition with applications"* **Sensors *25***(5): 1410. DOI: 10.3390/s25051410.

[15] K. Ashutosh and K. Grauman. "Learning skill-attributes for transferable assessment in video". In: *Advances in Neural Information Processing Systems*. ***38***. 2026, 160403–160430. DOI: 10.48550/arXiv.2511.13993.