IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 42, No. 2, April, 2025, pp. 729-736
Journal homepage: http://iieta.org/journals/ts

Fine-Grained Action Understanding in Instructional Sports Videos via a Hierarchical Spatiotemporal Pyramid Network

Songjiao Wu[1], Yuan Wang[2], Liping Wang[3*]

[1] School of Physical Education, Chongqing University of Arts and Sciences, Chongqing 402160, China
[2] Shijiazhuang College of Applied Technology, Shijiazhuang 050081, China
[3] Shijiazhuang Information Engineering Vocational College, Shijiazhuang 052161, China

Corresponding Author Email: 18831111916@139.com

https://doi.org/10.18280/ts.420211

Received: 12 October 2024
Revised: 27 February 2025
Accepted: 12 March 2025
Available online: 30 April 2025

Keywords:
fine-grained action understanding, instructional sports video, spatiotemporal pyramid network, multi-scale prediction algorithm, temporal scale, action prediction

ABSTRACT

With the digital transformation of sports education and athletic training, the automated analysis and understanding of instructional sports videos have emerged as critical areas of research. Fine-grained action understanding models play an increasingly significant role in this context, as they are designed to accurately extract and analyze detailed motion information. Traditional approaches to action recognition have primarily relied on single-scale feature extraction, which has proven inadequate for handling complex spatiotemporal information, especially in scenarios characterized by high variability and rapid motion transitions. These limitations often result in reduced accuracy and poor real-time performance. In recent years, multi-scale network models have been explored to enhance video analysis capabilities; however, challenges remain in balancing computational efficiency with precision. To address these shortcomings, a fine-grained action understanding model based on a hierarchical spatiotemporal pyramid network was proposed in this study. By constructing a multi-scale spatiotemporal pyramid prediction algorithm, this model can improve the extraction of spatiotemporal feature points of sports actions. In addition, by incorporating a temporal scale-based fine-grained action prediction algorithm, the model can capture intricate details within instructional sports videos accurately. By optimizing dynamic spatiotemporal characteristics and temporal dependencies, this study achieves improved accuracy and real-time performance in the prediction of fine-grained sports actions, offering a novel theoretical and technical foundation for the development of intelligent sports instruction systems.

1. INTRODUCTION

With the rapid advancement of digital technologies, the application scenarios of instructional sports videos have become increasingly diverse [1, 2], serving as essential tools for enhancing athletic skills and strengthening instructional effectiveness. Traditional methods of sports education have gradually shifted toward video-based analysis and intelligent instructional paradigms [3-7]. Accordingly, the accurate interpretation of actions depicted in video sequences [8] and the automated analysis and feedback of motion patterns [9] have attracted significant attention in both academia and industry. As a core technology within artificial intelligence-driven video analysis, fine-grained action understanding [10, 11] requires the extraction of detailed behavioral features from instructional videos [12]. A key challenge lies in the precise extraction and analysis of spatiotemporal feature points under complex and diverse motion conditions.

The task of action understanding in instructional videos typically involves processing large volumes of spatiotemporal information [13, 14]. Effectively extracting and interpreting this information across multiple temporal and spatial scales is critical for improving both the accuracy and real-time responsiveness of video analysis systems. Existing studies have predominantly focused on conventional deep learning models [15-17]. However, these models often struggle to simultaneously capture the dynamic spatiotemporal properties of motion and the intricacies of complex action patterns. In response to these limitations, a novel action understanding model based on a hierarchical spatiotemporal pyramid network was introduced in this study. This model enables the comprehensive extraction of spatiotemporal feature points from a multi-scale perspective while enhancing the precision of fine-grained action prediction. The proposed model offers a more efficient and accurate solution for analyzing instructional sports videos.

However, several significant limitations remain in existing studies on fine-grained action understanding. Wang et al. [18] performed feature extraction using a single-scale approach, which failed to account for the diversity of action expression and the variability of spatiotemporal scales, ultimately resulting in suboptimal prediction performance. Tavcar et al. [19] concentrated on the recognition of holistic actions and demonstrated limited capability in analyzing fine-grained motion components. This limitation was particularly evident in complex scenes, where the capture and precision of detailed

behaviors were frequently compromised. Furthermore, although multi-scale network models have improved action recognition performance to some extent, challenges related to computational efficiency and real-time processing persist when handling large-scale sports video datasets.

This study focuses on two core research components. First, a fast multi-scale spatiotemporal pyramid prediction algorithm was proposed for the computation of spatiotemporal feature points, in which a hierarchical spatiotemporal pyramid structure was constructed to enhance the processing capacity for multi-scale motion information. This design can significantly improve prediction accuracy and responsiveness. Second, a fine-grained action prediction algorithm for instructional sports videos based on temporal scale was introduced to investigate the temporal characteristics of different sports behaviors. This method can optimize temporal dependencies and the handling of fine-grained motion details throughout the action recognition process. Together, these contributions provide a novel framework and methodology for advancing fine-grained action understanding in instructional sports videos, offering important support for the application and development of intelligent instructional systems.

2. MULTI-SCALE SPATIOTEMPORAL PYRAMID PREDICTION ALGORITHM FOR COMPUTING ACTION SPATIOTEMPORAL FEATURE POINTS

In the context of fine-grained action understanding in instructional sports videos, the motion patterns captured often exhibit high complexity and variability. This necessitates a model capable of precisely identifying and analyzing subtle motion details. Fine-grained action understanding requires the extraction of information across multiple temporal and spatial scales. By adopting a hierarchical spatiotemporal pyramid network architecture, features at various scales can be computed through a layered strategy, thereby enhancing the capacity for recognizing diverse motion behaviors. However, such an approach is often accompanied by high computational overhead and limited processing speed. Especially when real-time performance is required, the computational efficiency of such an approach typically fails to meet practical demands, making it difficult for the model to run efficiently in dynamic and complex instructional sports environments. To overcome this bottleneck, a feature prediction algorithm based on inter-scale layered image structures was introduced as a solution to improve real-time performance. This algorithm effectively preserves computational accuracy while accelerating feature computation across scales, thereby offering a more efficient and practical framework for fine-grained action understanding in instructional sports videos. The proposed methodology can be introduced in two key parts: (a) the principle of a fast spatiotemporal pyramid hierarchical computation strategy for the multi-scale features of fine-grained actions in instructional sports videos; (b) a fine-grained action understanding algorithm based on temporal scale for those instructional sports videos.

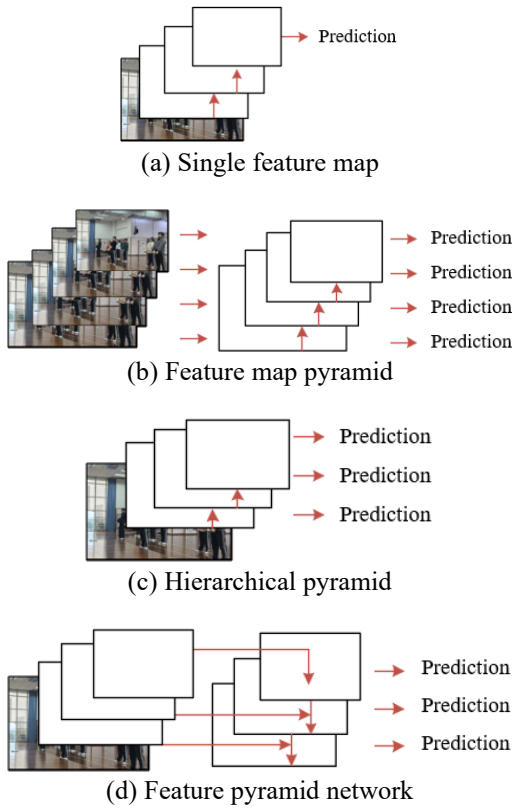
To achieve accurate analysis and interpretation of athletes' movements and postures, precise feature extraction from video frames is essential. In this process, the selection of image scales plays a decisive role in the successful extraction of fine-grained motion features. To effectively capture action details in instructional sports videos, multi-scale images (U_δ) were generated via upsampling and downsampling. In addition, their spatiotemporal feature points were extracted using the Hessian matrix. Let $h(a,b,s;\delta_u;\pi_u)=1/[(2\tau)^3\delta^4{}_u\pi^2{}_u]^{1/2}\times EXP(1(a^2+b^2)/2\delta^2-s^2/2\pi^2{}_u)$ and $M_{aa}=\partial/\partial a^2 h(\delta_u,\pi_u)\times I(a,b,s)$, and the Hessian matrix composed of second-order partial derivatives can be expressed as follows:

$$G\left(h;\delta_u^2;\pi_u^2\right)=\begin{pmatrix} M_{aa} & M_{ab} & M_{as} \\ M_{ba} & M_{bb} & M_{bs} \\ M_{sa} & M_{sb} & M_{ss} \end{pmatrix} \tag{1}$$

The Hessian matrix, as a scale-invariant feature extraction tool, enables the accurate and stable localization of salient feature points in images. Given the significant temporal and spatial variability of actions in instructional sports videos, traditional feature extraction methods often fail to adapt effectively to such variations. In contrast, the Hessian matrix ensures consistent detection of critical motion features across multiple scales, and it provides robust support for the precise localization and extraction of important spatiotemporal feature points, particularly in fine-grained action recognition. To further simplify the computation and to ensure that the Gaussian kernel accurately selects appropriate spatial and temporal scales, denoted as δ_0 and π_0, a γ-normalization process was applied to the Hessian matrix, enhancing both computational efficiency and accuracy. The absolute value of the determinant of the Hessian matrix is expressed as:

$$\begin{aligned} T &= \left|DE\left(G^{NORM}\right)\right| = M_{aa}^{\gamma-NORM} M_{bb}^{\gamma-NORM} M_{ss}^{\gamma-NORM} \\ &= \delta^{2o}\pi^{2w} M_{aa} M_{bb} M_{ss} = \delta^5 \pi^{5/2} M_{aa} M_{bb} M_{ss} \end{aligned} \tag{2}$$

where, o=5/2, and w=5/4.



(a) Single feature map

(b) Feature map pyramid

(c) Hierarchical pyramid

(d) Feature pyramid network

Figure 1. Various strategies for multi-scale feature extraction and fusion

Figure 1 illustrates different strategies for multi-scale feature extraction and fusion. Traditional spatiotemporal pyramid approaches perform repeated resampling of image data across various scales. The input images are divided into multiple hierarchical groups, each further subdivided into finer-scale layers. While this method enables the extraction of a rich set of spatiotemporal features, it incurs a substantial computational burden. In particular, repeated sampling at each scale results in significant inefficiencies. In real-world instructional sports video analysis, systems are required to respond rapidly and deliver real-time feedback. This is especially critical in live instructional settings, where both instructors and learners depend on the immediate analysis of motion details. The traditional multi-scale spatiotemporal pyramid strategy faces significant real-time issues in practical applications, making it difficult to meet the dual requirements of real-time performance and accuracy.

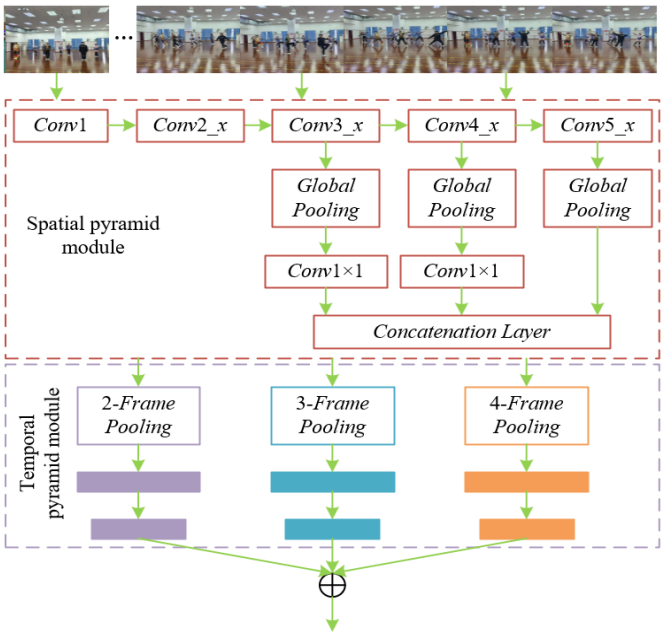
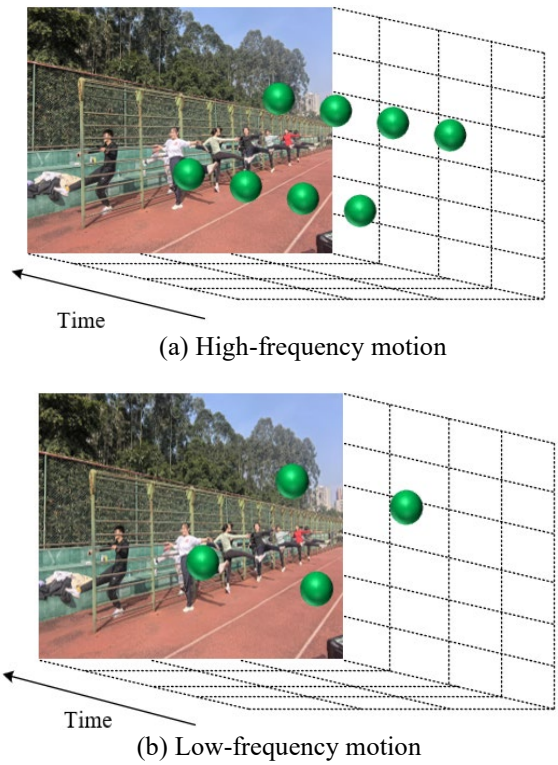


Figure 2. Schematic diagram of the fast multi-scale spatiotemporal pyramid prediction model

To overcome the aforementioned limitations, a fast multi-scale spatiotemporal pyramid prediction algorithm was introduced in this study to accelerate the computation of spatiotemporal features without compromising accuracy. The structure is illustrated in Figure 2. Through this algorithm, redundant computational steps were eliminated, significantly reducing the processing time required per video frame. In this scheme, video images were initially divided into multiple scale groups following conventional methods. However, unlike traditional approaches, only a single downsampling operation was performed for each group, reducing the repetitive calculation of each scale. Specifically, images at scales of 1/2 and 1/4 were computed, and the determinant of the Hessian matrix, denoted as $|DET(G^{NORM})|$, was utilized to extract the values of feature points. This approach substantially reduces computational complexity by eliminating the need for repeated operations across all layers, effectively accelerating the extraction of spatiotemporal feature points. To further enhance efficiency, a feature point computation strategy based on the prediction algorithm was adopted in this study. After calculating the feature point values T at the 1/2 and 1/4 scales, these known values were then used to predict the feature point values at adjacent scales. The prediction algorithm infers feature point values of other scales based on the relationship between neighboring scales, avoiding the need for computation at every individual scale.

3. FINE-GRAINED ACTION UNDERSTANDING IN INSTRUCTIONAL SPORTS VIDEOS BASED ON TEMPORAL SCALE

In the context of fine-grained action understanding for instructional sports videos, athlete movements often exhibit distinct characteristics across different temporal and spatial scales. To analyze these complex spatiotemporal patterns, a temporal scale-based prediction algorithm was employed in this study to efficiently extract spatiotemporal feature points under varying motion frequencies. The core principle of this algorithm lies in analyzing the relationship between the temporal and spatial scales of each spatiotemporal feature point, thereby intelligently predicting and computing the positions of feature points across multiple temporal scales. For instance, fast-paced actions such as shooting or sprinting are typically associated with smaller temporal scales, whereas slower actions such as standing or preparing involve larger temporal scales. In the case of a waving gesture, if the frequency of waving is high, the corresponding spatiotemporal feature points are concentrated at smaller temporal scales. These shorter temporal scales are capable of capturing rapid motion transitions, enabling more precise representation of hand movement details. Conversely, when the waving frequency is lower, the associated feature points shift to larger temporal scales, reflecting slower motion transitions and producing feature point calculations characterized by extended temporal scales. Figure 3 illustrates the computation of action-related spatiotemporal feature points under varying motion frequencies.



(a) High-frequency motion

(b) Low-frequency motion

Figure 3. Illustration of spatiotemporal feature point computation under varying motion frequencies

In fine-grained action understanding for instructional sports videos, the concept of scale space was utilized to analyze and recognize complex motion sequences through multi-level smoothing operations. Within this context, the scale space can be defined as a family of image smoothing operators S_π, representing the degree of processing applied to each frame of the video. By progressively increasing the scale π, the degree of smoothing is intensified, allowing low-frequency noise to be suppressed while enhancing the visibility of fine-grained action features present in the video. Each smoothing operator S_π in the scale space can be understood as applying different levels of blurring to incrementally capture and refine variations in motion, which is critical for the accurate recognition of actions in instructional settings. As the value of π varies, a continuous transition from localized motion to more global action patterns is exhibited, enabling effective differentiation and recognition of fine-grained actions across scales—such as the nuanced progression of actions like shooting a basketball or kicking a soccer ball.

Figure 4. Temporal reasoning illustration of the hierarchical spatiotemporal pyramid network

In the hierarchical spatiotemporal pyramid network architecture adopted in the model, the use of scale space further enhanced the spatiotemporal pyramid structure of the video. By progressively extracting features at varying scales from video data, fine-grained action recognition was achieved. More specifically, the scale space was employed to construct multi-layered, cross-scale feature extraction pathways within the network, enabling the capture of subtle motion variations across spatiotemporal scales. Under the influence of this multi-scale framework, the network is capable of performing scale transitions using the operator $S_{\pi+g,\pi}$, facilitating precise modeling of complex behaviors. Through the progressive refinement enabled by the scale space, the model was equipped to interpret and recognize a broad range of instructional sports actions in fine detail, including action initiation, transitional movements, and termination. Figure 4 illustrates the temporal reasoning process within the hierarchical spatiotemporal pyramid network.

In this study, the temporal scale configuration follows a recursive principle, whereby the spatiotemporal pyramid structure of the video is divided into v layers, with each layer's temporal scale defined as a multiple (π, 2π, ..., $v\pi$) of a base unit π. This design enables the effective capture of hierarchical temporal and spatial variations within the video. At different scales, features can be extracted layer by layer, and the mapping between scales π and $\pi+g$ through $S_{\pi+v}$ iterations yields fine spatiotemporal variation features. As a result, both low-frequency global motions and high-frequency local details are precisely modeled at each layer of the model. Specifically, the recursive structure of the scale space is redefined as follows:

$$\begin{aligned} M_1 &= h(a,b,s;\delta,\pi)*U_0 \\ M_2 &= h(a,b,s;\delta,2\pi)*M_1 \\ &\cdots \\ M_v &= h(a,b,s;\delta,v\pi)*M_{v-1} \end{aligned} \tag{3}$$

Finally, the temporal scale-based prediction algorithm further optimizes the action understanding model by calculating the ratio of the determinants of the Hessian matrix across different spatial and temporal scales:

$$\frac{T_1}{T_u} = \frac{\delta_1^5 \pi_1^{5/2} M_{aa} M_{bb} M_{ss}}{\delta_u^5 \pi_u^{5/2} M_{aa} M_{bb} M_{ss}} = \frac{\delta_1^4 M_{aa} M_{bb}}{\delta_u^4 M_{aa} M_{bb}} \cdot \frac{\delta_1 \pi_1^{5/2} M_{ss}}{\delta_u \pi_u^{5/2} M_{ss}} \tag{4}$$

Based on experimental results, the following relationship was established:

$$\frac{d_{FpG}(\delta_1)}{d_{FpG}(\delta_u)} = \frac{\delta_1^4 M_{aa} M_{bb}}{\delta_u^4 M_{aa} M_{bb}} = 2\left(\frac{\delta_1}{\delta_u}\right)^{0.72} \tag{5}$$

Therefore, it can be concluded that:

$$M_{ss} = \frac{\partial^2 h}{\partial a^2} * U = \frac{s^2 - 1}{\pi_u^2} \cdot h(\pi_u) * U \tag{6}$$

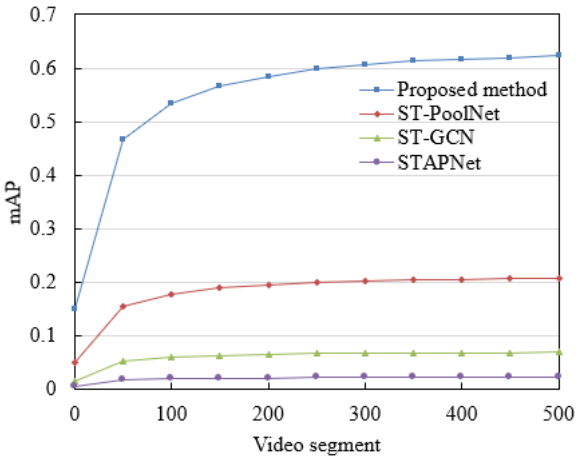
$$\begin{aligned} \frac{T_1}{T_u} &= 2\left(\frac{\delta_1}{\delta_u}\right)^{0.72} \frac{\delta_1 \pi_1^{5/2}}{\sigma_u \pi_u^{5/2}} \frac{\frac{s^2-1}{\pi_1^2} \cdot h(\pi) * U}{\frac{s^2-1}{\pi_1^2} \cdot h(u\pi) * U} \\ &= 2\left(\frac{\delta_1}{\delta_u}\right)^{1.72} \left(\frac{\pi_1}{u\pi}\right)^{1/2} \frac{h(\pi) * U}{h(u\pi) * U} \end{aligned} \tag{7}$$

Since the temporal scale is defined as a multiple of π, it follows that $\pi_u = u\pi$.

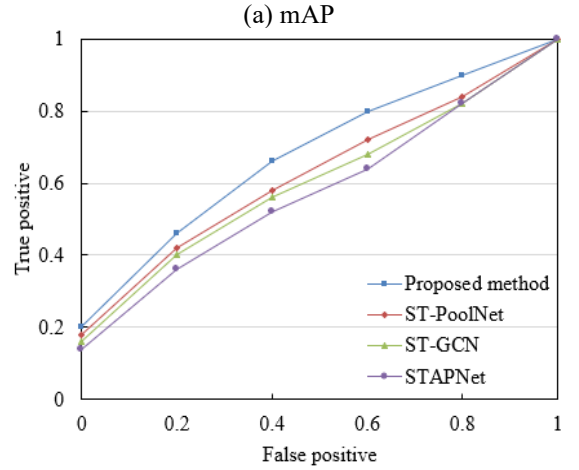
4. EXPERIMENTAL RESULTS AND ANALYSIS

The experimental results presented in Figure 5 reveal substantial differences in performance among the models employing different pyramid prediction algorithms, particularly in multi-scale action prediction and Receiver

Operating Characteristic (ROC) curve behavior across video segments. In the performance comparison based on mean Average Precision (mAP), the proposed method exhibited the highest predictive accuracy. As the duration of the video segments increased, prediction performance improved steadily, ultimately reaching an mAP value of 0.623. In contrast, ST-PoolNet and ST-GCN demonstrated inferior results. ST-PoolNet maintained consistently low performance, achieving a maximum mAP of only 0.2076 over time. ST-GCN performed moderately better, reaching 0.0692, yet remained far below the performance of the proposed approach. The weakest performance was observed in STAPNet, which attained a peak value of merely 0.0231 across multiple temporal scales. These results indicate that the proposed hierarchical spatiotemporal pyramid prediction algorithm provides a clear advantage in fine-grained action understanding for video segments, significantly enhancing the model's multi-scale information processing ability and predictive accuracy.



(a) mAP



(b) ROC curve

Figure 5. Comparative performance of different pyramid prediction algorithms

From the ROC curve analysis, the proposed method also outperformed the comparative models across all threshold values. The curve approached the ideal value of 1.0, with an accuracy of 0.9 achieved at a threshold of 0.8. These findings suggest that the proposed algorithm is highly effective in capturing spatiotemporal features in video segments and excels in complex action recognition tasks. By contrast, ST-PoolNet, ST-GCN, and STAPNet exhibited notably weaker performance in the ROC evaluation. Although all three models approached a maximum near 1.0, significant performance disparities were observed across the threshold range of 0.2 to 0.8, where the proposed method consistently maintained superior accuracy. In summary, the multi-scale spatiotemporal pyramid prediction algorithm presented in this study effectively integrates spatiotemporal features, optimizes temporal dependencies in actions and the processing of motion detail, thereby substantially increasing the precision of fine-grained action prediction and understanding in instructional sports videos.

Table 1. Prediction results for fine-grained actions in different instructional sports videos

Action Type	Training Set Accuracy	Validation Set Accuracy
Demonstration	67.25%	85.32%
Error correction	71.23%	91.25%
Imitative practice	72.59%	91.36%
Interactive questioning	73.65%	92.35%
Performance display	71.79%	90.51%

The prediction results presented in Table 1 indicate that the proposed hierarchical spatiotemporal pyramid network architecture demonstrates superior performance in recognizing fine-grained actions across various types of instructional sports videos. The model achieved consistently high accuracy on both the training and validation sets for each category of instructional behavior. Specifically, an accuracy of 67.25% was achieved for "demonstration" in the training set and 85.32% in the validation set. For "error correction," the training accuracy reached 71.23%, while the validation accuracy improved to 91.25%. "Imitative practice" resulted in training and validation accuracies of 72.59% and 91.36%, respectively. In the case of "interactive questioning," accuracies of 73.65% and 92.35% were recorded, and for "performance display," the model achieved 71.79% and 90.51% on the training and validation sets, respectively. These results demonstrate that the proposed method is capable of effectively processing fine-grained instructional actions across diverse video scenarios. In particular, the high validation accuracies reflect the model's strong generalization ability.

Table 2. Experimental results of the proposed model on different types of datasets

Dataset	Training Set Accuracy	Validation Set Accuracy
MultiSports	91.25%	97.56%
FineDiving	91.27%	97.52%
FineSports	91.36%	96.32%
Average	91.58%	97.54%

As shown in Table 2, the proposed hierarchical spatiotemporal pyramid network architecture achieved consistently high performance across multiple dataset types. High accuracy and stability were observed on both the training and validation sets. For example, on the MultiSports dataset, accuracies of 91.25% and 97.56% were recorded for the training and validation sets, respectively. On the FineDiving dataset, the training and validation accuracies reached 91.27% and 97.52%, respectively. Similarly, the FineSports dataset yielded accuracies of 91.36% (training) and 96.32% (validation). These results demonstrate that the proposed model consistently achieved strong predictive performance

across varied datasets. In particular, the validation set accuracies were observed to be consistently higher than those of the training set, indicating strong generalization capability during the training process. Overall, the model achieved average accuracies of 91.58% on the training sets and 97.54% on the validation sets, further validating the effectiveness of the proposed method.

Table 3. Comparative experimental results of different action prediction networks

Network Model	*MultiSports*	*FineDiving*	*FineSports*
FineGym	81.2%	82.4%	81.8%
SlowFast	83.6%	84.2%	84.5%
Temporal Segment Networks	84.5%	85.9%	83.9%
Temporal Excitation and Aggregation	85.6%	86.7%	84.3%
Temporal Difference Networks	86.9%	87.4%	86.2%
MotionLLM	87.5%	88.7%	86.4%
Proposed model	91.6%	90.4%	90.7%

As shown in Table 3, the hierarchical spatiotemporal pyramid network architecture proposed in this study demonstrated a clear performance advantage when compared with other action prediction networks across all evaluated datasets. Notably, the proposed model consistently achieved the highest accuracy scores on three datasets. For example, on the MultiSports dataset, an accuracy of 91.6% was attained, surpassing MotionLLM (87.5%) and Temporal Difference Networks (86.9%). On the FineDiving dataset, the proposed model achieved 90.4%, outperforming MotionLLM (88.7%) and Temporal Excitation and Aggregation (86.7%). Similarly, on the FineSports dataset, the proposed approach reached 90.7%, exceeding the performance of MotionLLM (86.4%) and SlowFast (84.5%). These results confirm the superior effectiveness of the proposed hierarchical spatiotemporal pyramid network in fine-grained action prediction across diverse instructional sports video datasets.

As shown in the confusion matrices presented in Figure 6, in the MultiSports dataset, the classification accuracy for the "demonstration" category reached 1.00, indicating perfect recognition. However, 25% of the "error correction" samples were misclassified as "imitative practice," and 50% of the "imitative practice" samples were misclassified as "interactive questioning." For the FineDiving dataset, both "demonstration" and "performance display" achieved perfect accuracy (1.00). In contrast, 25% of the "error correction" samples were misclassified as "performance display," and 25% of the "imitative practice" samples were assigned to "interactive questioning." In the FineSports dataset, the accuracy for "demonstration" again reached 1.00. However, 75% of the "error correction" samples were misclassified as "interactive questioning." Additionally, 25% of the "imitative practice" samples were assigned to "demonstration," and 25% of "interactive questioning" samples were misclassified as "performance display."

Overall, the model demonstrated exceptional accuracy in recognizing the "demonstration" and "performance display" categories across all datasets, indicating strong discriminative ability for these action types. However, misclassifications were more common among the "error correction," "imitative practice," and "interactive questioning" categories. This may be attributed to the intrinsic similarity in their features, which complicates the model's ability to differentiate between them. The observed variation in misclassification patterns across datasets also reflects the influence of dataset-specific characteristics on recognition performance. These findings suggest that further refinement is needed to enhance the model's ability to distinguish between closely related instructional behaviors, thereby improving the overall performance of the model in fine-grained action understanding in instructional sports videos.

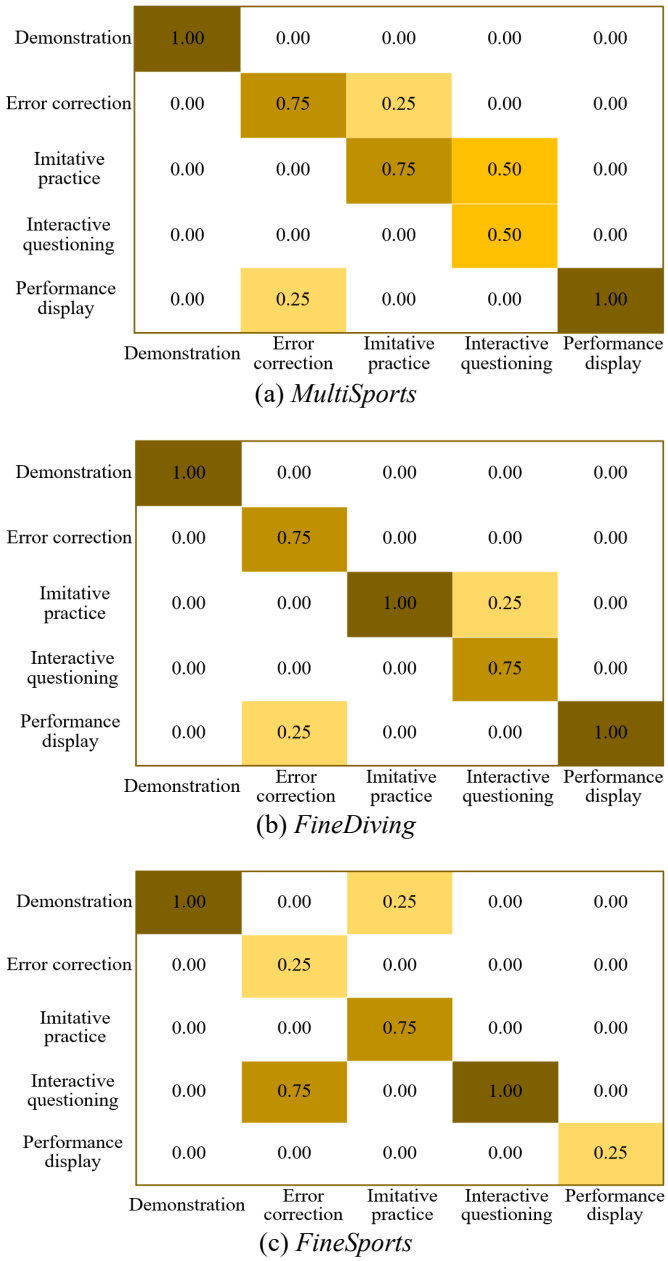


Figure 6. Confusion matrices of the proposed model across different dataset types

5. CONCLUSION

The proposed hierarchical spatiotemporal pyramid network architecture has significantly advanced the task of fine-grained action understanding in instructional sports videos through two core innovations. First, a fast multi-scale spatiotemporal pyramid prediction algorithm for computing action-related spatiotemporal feature points was introduced. This approach

enhanced the model's ability to process multi-scale motion information, thereby improving both prediction accuracy and real-time performance. Second, a temporal scale-based fine-grained action prediction algorithm was proposed, enabling the detailed analysis of temporal features across different types of sports actions and optimizing both temporal dependencies and fine-detail processing in the recognition process. The integration of these two innovations resulted in substantial improvements in the accuracy and interpretability of action recognition within instructional sports video contexts. Furthermore, this framework provides new insights for the development of intelligent instructional systems, offering considerable potential for practical deployment and further advancement.

Despite the promising results achieved across multiple datasets, several limitations remain. First, the current model exhibits relatively high training and inference latency, particularly when deployed on large-scale video datasets. This presents challenges for real-time applications due to the demand for substantial computational resources. Second, although the proposed approach effectively processes multi-scale spatiotemporal features, performance may still degrade in highly complex and dynamically evolving behavioral scenes. The ability in handling long-duration sequences or actions with elevated structural complexity should be further enhanced. Future research efforts should focus on improving computational efficiency and enhancing the model's capacity to process spatiotemporal information in challenging scenarios and exploring the lightweight models tailored for real-time applications. Additionally, as the diversity of instructional sports videos continues to grow, expanding the generalization capabilities of the model will be essential to ensure reliable performance across a wider range of behavioral categories.

REFERENCES

[1] Zhu, X., Zhang, Z. (2023). Precise recommendation algorithm for online sports video teaching resources. EAI Endorsed Transactions on Scalable Information Systems, 10(2): e11. https://doi.org/10.4108/eetsis.v10i1.2578

[2] Mącznik, A.K., Schneiders, A.G., Athens, J., Sullivan, S.J. (2018). The development of an instructional video for the teaching of acupressure for pain management in acute musculoskeletal injuries: A knowledge translation study. Physical Therapy in Sport, 29: 34-42. https://doi.org/10.1016/j.ptsp.2017.10.005

[3] Gómez, R.S. (2020). La enseñanza de las actividades físicas de incertidumbre ambiental en Educación Física: en busca de una performance inteligente para los jugadores de la naturaleza. Ágora para la Educación Física y el Deporte, 22: 296-319. https://doi.org/10.24197/aefd.0.2020.296-319

[4] SueSee, B., Pill, S., Edwards, K. (2018). Interrogating assumptions of a curriculum: queensland senior physical education syllabus. Physical Educator, 75(5): 850-879. https://doi.org/10.18666/TPE-2018-V75-I5-8283

[5] Zeng, B., Zhao, J., Wen, S. (2023). A textual and visual features-jointly driven hybrid intelligent system for digital physical education teaching quality evaluation. Mathematical Biosciences and Engineering, 20(8): 13581-13601. https://doi.org/10.3934/mbe.2023606

[6] Li, C., Liu, B., Kim, K. (2023). Intelligent unsupervised learning method of physical education image resources based on genetic algorithm. Neural Computing and Applications, 35(6): 4225-4242. https://doi.org/10.1007/s00521-022-07021-x

[7] Xu, Q., Yin, J. (2021). Application of random forest algorithm in physical education. Scientific Programming, 2021(1): 1996904. https://doi.org/10.1155/2021/1996904

[8] Gayathri, T., Mamatha, H.R. (2023). How to improve video analytics with action recognition: A survey. ACM Computing Surveys, 57(1): 9. https://doi.org/10.1145/3679011

[9] Nguyen, L.Q., Choi, J., Dang, L.M., Moon, H. (2024). Background debiased class incremental learning for video action recognition. Image and Vision Computing, 151: 105295. https://doi.org/10.1016/j.imavis.2024.105295

[10] Yıldız, M., Keskin, H.K., Oyucu, S., Hartman, D.K., Temur, M., Aydoğmuş, M. (2025). Can artificial intelligence identify reading fluency and level? comparison of human and machine performance. Reading & Writing Quarterly, 41(1): 66-83. https://doi.org/10.1080/10573569.2024.2345593

[11] Sharma, V., Gupta, M., Pandey, A.K., Mishra, D., Kumar, A. (2022). A review of deep learning-based human activity recognition on benchmark video datasets. Applied Artificial Intelligence, 36(1): 2093705. https://doi.org/10.1080/08839514.2022.2093705

[12] Fernández-Ramírez, J., Álvarez-Meza, A., Pereira, E. M., Orozco-Gutiérrez, A., Castellanos-Domínguez, G. (2020). Video-based social behavior recognition based on kernel relevance analysis. The Visual Computer, 36(8): 1535-1547. https://doi.org/10.1007/s00371-019-01754-y

[13] Naylor, A., Spence, S.E., Poed, S. (2019). Using video modelling to teach expected behaviours to primary students. Support for Learning, 34(4): 389-403. https://doi.org/10.1111/1467-9604.12274

[14] Yoon, H.Y., Kang, S., Kim, S. (2024). A non-verbal teaching behaviour analysis for improving pointing out gestures: The case of asynchronous video lecture analysis using deep learning. Journal of Computer Assisted Learning, 40(3): 1006-1018. https://doi.org/10.1111/jcal.12933

[15] van Dam, E.A., Noldus, L.P., van Gerven, M.A. (2020). Deep learning improves automated rodent behavior recognition within a specific experimental setup. Journal of Neuroscience Methods, 332: 108536. https://doi.org/10.1016/j.jneumeth.2019.108536

[16] Teterja, D., Garcia-Rodriguez, J., Azorin-Lopez, J., Sebastian-Gonzalez, E., Net al. (2024). A Video Mosaicing-Based Sensing Method for Chicken Behavior Recognition on Edge Computing Devices. Sensors, 24(11): 3409. https://doi.org/10.3390/s24113409

[17] Rezaei, F., Yazdi, M. (2021). Real-time crowd behavior recognition in surveillance videos based on deep learning methods. Journal of Real-Time Image Processing, 18(5): 1669-1679. https://doi.org/10.1007/s11554-021-01116-9

[18] Wang, X., Song, Y., Hou, F., Zhang, M., Richardson, A. G., Lucas, T.H., Van der Spiegel, J. (2022). Design of a real-time movement decomposition-based rodent tracker and behavioral analyzer based on FPGA. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 30(9): 1133-1143. https://doi.org/10.1109/TVLSI.2022.3168783

[19] Tavcar, A., Kuznar, D., Gams, M. (2017). Hybrid multi-agent strategy discovering algorithm for human behavior. Expert Systems with Applications, 71: 370-382. https://doi.org/10.1016/j.eswa.2016.11.036