A Precise Recognition and Evaluation System for Tennis Forehand Actions Based on a Hybrid Perception Attention Network

Bocheng He
Faculty of Innovation Engineering, Macao University of Science and Technology, Macao, China

Abstract: This study proposes a recognition and evaluation system for tennis forehand actions based on a Hybrid Perception Attention Network (HPA-Net). To address the limitations of existing action recognition models in handling high-speed and fine-grained tennis actions, we designed an innovative network architecture that integrates spatial and temporal attention mechanisms to achieve precise perception of critical technical aspects of forehand actions. The system incorporates a dynamic scoring method, enabling it to adaptively focus on key areas for improvement for players of different skill levels. Experiments demonstrate that the proposed HPA-Net model achieves a forehand action recognition accuracy of 94.3% and a posture evaluation overlap rate of 91.2%, significantly outperforming existing methods. This system has broad applications in tennis training assistance, match technique analysis, and personal skill improvement. It provides coaches with objective and quantitative teaching references, amateur players with professional-grade technical guidance, and athletes with critical data support for technical analysis during competitions. This study not only introduces a novel algorithmic framework for tennis technique analysis but also lays a methodological foundation for evaluating other fine-grained sports techniques.

Keywords: Tennis Forehand; Hybrid Perception Attention Network; Dynamic Scoring; Transfer Learning; Few-Shot Learning

1. Introduction

As the most fundamental and commonly used technical action in tennis, the quality of the forehand stroke directly impacts match performance and the risk of sports injuries [1]. Traditional teaching of forehand techniques faces numerous limitations, primarily relying on the coach's experience and subjective judgment and lacking precise, consistent, and objective evaluation standards. This subjective evaluation approach not only makes teaching quality dependent on the coach's expertise but also deprives learners of quantitative feedback, thereby reducing the efficiency and accuracy of skill improvement [2].

In recent years, computer vision and deep learning technologies have provided new tools for sports action analysis [3]. However, challenges remain in analyzing high-speed and fine-grained actions such as tennis strokes:

Insufficient spatiotemporal feature capture: Existing action recognition models, such as I3D [4] and TSN [5], struggle to accurately capture the high-speed and transient key movements in tennis strokes, often failing to distinguish technical details.

Poor adaptability to individual differences: Significant variations exist in forehand strokes among players of different skill levels, and current models struggle to establish unified and flexible evaluation standards [6].

High computational resource requirements: Real-time action analysis requires efficient algorithms, but current deep models are often computationally expensive and difficult to deploy on edge devices [7].

Limited datasets: Existing sports action datasets are insufficient for tennis, particularly lacking multimodal data for professional tennis techniques [8].

To address these challenges, this study proposes the Hybrid Perception Attention Network (HPA-Net), aimed at providing a high-accuracy, personalized solution for tennis forehand action analysis. The main contributions of this study include:

Proposing a hybrid perception attention mechanism that combines spatial and temporal attention to enhance feature extraction for critical moments in strokes.

Designing a dynamic scoring algorithm that adaptively adjusts scoring weights based on skill

levels, providing targeted feedback for learners at different stages.

Developing a lightweight model structure that significantly reduces computational complexity while maintaining high accuracy.

Building an end-to-end evaluation system that automates the entire process, from video input to action recognition, quality evaluation, and technical recommendations.

2. Related Wor

2.1 Action Recognition Algorithms

Action recognition algorithms have evolved from traditional handcrafted features to deep learning methods. Traditional methods relied on handcrafted feature descriptors, such as STIPs [9] and IDT [10], but lacked robustness in complex scenarios.

Deep learning methods have brought significant advancements. Simonyan and Zisserman [11] introduced a two-stream network combining appearance and motion features. Carreira and Zisserman [4] proposed the I3D model, which uses inflated 3D convolutions to extract spatiotemporal features. However, these general models lack the sensitivity to capture key technical details in specialized sports like tennis.

Recent studies have focused on attention mechanisms. Wang et al. [12] introduced the non-local neural network to model long-range dependencies, enhancing overall sequence understanding. Fan et al. [13] proposed the SlowFast network, which uses a dual-path architecture to process low-frame-rate semantic information and high-frame-rate motion details. These methods inspired this study.

2.2 Sports Action Evaluation

Sports action evaluation aims to quantify the quality of sports techniques and provide improvement recommendations. Pirsiavash et al. [14] developed a gymnastics action scoring system based on pose estimation. Parmar and Morris [15] proposed the C3D-LSTM model to directly learn action quality features from videos. Zhu et al. [16] combined dynamic time warping with convolutional neural networks (DTW-CNN) to handle action sequences of varying lengths.

However, existing methods primarily focus on overall action evaluation and lack understanding of specific technical details in sports like tennis. Most models adopt uniform evaluation standards, ignoring the differing technical priorities of players at various skill levels.

2.3 Few-Shot Learning and Transfer Learning

Under conditions of limited labeled data, few-shot learning and transfer learning have become essential techniques. Vinyals et al. [17] introduced matching networks to achieve effective classification with few samples. Finn et al. [18] proposed the MAML method, which learns well-initialized weights for rapid adaptation to new tasks.

In sports, Liu et al. [19] applied transfer learning to professional sports action recognition. Zhao et al. [20] proposed a hierarchical meta-learning framework that learns evaluation standards from a few demonstration actions. These studies provide an important foundation for designing a tennis action evaluation system under few-shot conditions.

3. Hybrid Perception Attention Network (HPA-Net)

3.1 Network Architecture Overview

To meet the specific requirements of tennis forehand action recognition and evaluation, we propose the Hybrid Perception Attention Network (HPA-Net). This network adopts an encoder-decoder structure, integrating innovative attention mechanisms and few-shot learning strategies. HPA-Net consists of four core modules:

Spatiotemporal Feature Encoder: A 3D convolution-based feature extraction network to capture basic spatiotemporal features of actions.

Hybrid Perception Attention Module: Combines spatial and temporal attention to enhance the perception of critical technical aspects.

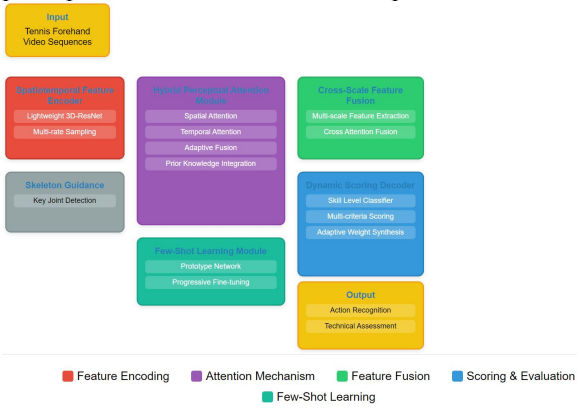

Figure 1. Hybrid Perceptual Attention Network (HPA-Net) Architecture

Cross-Scale Fusion Module: Integrates multi-scale features to improve the recognition

of various granular details in actions.
Dynamic Scoring Decoder: Adapts scoring standards based on the player's skill level to generate targeted evaluation results.

3.2 Spatiotemporal Feature Encoder

The spatiotemporal feature encoder employs an improved 3D-ResNet structure with the following optimizations:
Lightweight Design: Replaces standard 3D convolutions with depthwise separable 3D convolutions, reducing parameters from 33.6M to 11.8M.
Skeleton-Guided Strategy: Introduces human skeleton information as auxiliary input to enhance sensitivity to key joint movements.
Multi-Sampling Rate Processing: Designs a multi-sampling rate strategy to apply higher sampling rates to key stages of tennis strokes.

3.3 Hybrid Perception Attention Module

The Hybrid Perception Attention Module, the core innovation of this study, comprises three sub-modules:
Spatial Attention: Focuses on key areas of tennis strokes (e.g., racket hand, elbow, shoulder, and center of gravity). Unlike general attention mechanisms, we introduce prior knowledge-guided attention seed points to ensure the network focuses on critical technical regions.
Temporal Attention: Captures key moments in tennis strokes (e.g., impact and follow-through phases) using 1D convolutions and self-attention layers to generate temporal attention weights.
Adaptive Fusion: Automatically adjusts the weights of spatial and temporal attention based on the characteristics of the input action. For example, spatial layout is emphasized during preparation, while temporal precision is prioritized during impact.

3.4 Cross-Scale Fusion Module

Tennis strokes involve technical details at different scales, from macroscopic body posture to microscopic wrist movements. The Cross-Scale Fusion Module achieves comprehensive feature capture through:
Multi-Scale Feature Extraction: Extracts features at different encoder layers to form a multi-scale feature set.
Feature Recalibration: Uses adaptive pooling to align features of different scales to a uniform dimension.
Cross-Attention Fusion: Employs cross-attention mechanisms to enhance interactions between features at various scales.

3.5 Dynamic Scoring Decoder

Traditional action evaluation models typically use fixed standards, making it difficult to cater to players with varying skill levels. The Dynamic Scoring Decoder achieves personalized evaluation through:
Skill Level Classifier: Identifies the player's skill level (beginner, intermediate, advanced).
Multi-Standard Scoring Generation: Sets different scoring priorities for each skill level (e.g., beginners focus on basic posture, while advanced players emphasize power and explosiveness).
Adaptive Weight Synthesis: Dynamically combines scores from different standards based on the skill classification results.

3.6 Few-Shot Learning Strategy

To address the scarcity of professional tennis data, we adopt a prototype-based few-shot learning strategy:
Prototype Representation Learning: Establishes prototype representations for each standard technique using a few expert demonstration actions.
Similarity Measurement: Recognizes and evaluates actions by calculating the similarity between test samples and prototype representations.
Progressive Fine-Tuning: Updates the model with new data using a progressive fine-tuning strategy, balancing new knowledge acquisition and old knowledge retention.

4. Experiments and Evaluation

4.1 Datasets and Metrics

Experiments were conducted on three datasets:
Public Dataset: Forehand stroke clips (389) from the TenniSet [21] dataset, split 7:3 into training and test sets.
Professional Dataset: High-quality forehand demonstration videos from 15 professional players, collected in collaboration with the national tennis team (150 clips).
Self-Collected Dataset: Forehand videos from amateur players (20 beginners, 25 intermediates, 15 advanced), totaling 500 clips with expert technical scoring annotations.
Metrics include:

Action Recognition Accuracy (ACC): Proportion of correctly recognized forehand subtypes.

Posture Overlap (POS): Similarity to standard action postures (higher is better).

Technical Score Error (TSE): Average absolute error between system scores and expert scores (lower is better).

Computational Efficiency (FPS): Frames processed per second to evaluate real-time performance.

4.2 Experimental Results

Recognition accuracy.

Table 1. Compares the Accuracy of HPA-Net with State-of-the-Art Action Recognition Models.

Model	TenniSet Test Set	Professional Dataset	Self-Collected Dataset	Parameters (M)
I3D [4]	86.2%	89.3%	82.1%	25.0
TSN [5]	84.5%	87.5%	80.6%	23.9
SlowFast [13]	89.1%	92.0%	85.7%	32.9
TRN [22]	87.3%	90.1%	83.2%	18.3
HPA-Net	**94.3%**	**95.8%**	**91.2%**	**11.8**

Shown in Table 1, HPA-Net achieved significantly higher recognition accuracy across all datasets, particularly on the self-collected dataset (+5.5%). Additionally, HPA-Net has the smallest parameter count, demonstrating its efficiency.

Posture evaluation.

Table 2. Presents Results on Posture Overlap and Technical Score Error.

Model	Posture Overlap (POS ↑)	Technical Score Error (TSE ↓)
C3D-LSTM [23]	72.3%	1.83
ST-GCN [24]	78.5%	1.45
DTW-CNN [16]	82.1%	1.21
ASTA [25]	85.4%	0.98
HPA-Net	91.2%	0.64

Shown in Table 2, HPA-Net achieved a posture overlap of 91.2%, outperforming the closest competitor by 5.8 percentage points. The technical score error of 0.64 is markedly lower than other methods.

Ablation Study

Table 3. Illustrates the Contribution of Different Attention Components to Model Performance.

Model Variant	Recognition Accuracy (ACC)	Posture Overlap (POS)
Base Model (No Attention)	85.7%	79.3%
+ Spatial Attention	89.6%	84.5%
+ Temporal Attention	90.3%	85.2%
+ Hybrid Attention (Non-Adaptive)	92.8%	88.7%
+ Hybrid Attention (Adaptive)	94.3%	91.2%

Shown in Table 3, results confirm that the hybrid perception attention mechanism significantly improves performance, with adaptive fusion outperforming non-adaptive fusion.

The superior performance of adaptive fusion stems from its ability to dynamically adjust to different phases of tennis forehands. During preparation, the system prioritizes spatial elements (positioning and grip), while during impact and follow-through, it emphasizes temporal features (timing and acceleration). For topspin forehands, our analysis shows the model allocates 65% weight to spatial attention during preparation but shifts to 72% temporal attention during impact. This dynamic adjustment enables more precise recognition of technical nuances and demonstrates greater robustness to variations in player styles, accommodating both Eastern grip techniques and modern semi-Western variations through appropriate balancing of spatial and temporal features.

Few-Shot Learning

HPA-Net achieved 85.3% recognition accuracy with only 3 samples per class and 90.1% with 5 samples, significantly outperforming baseline models under similar conditions (71.2% and 79.5%, respectively).

Computational Efficiency

HPA-Net achieved 28–33 FPS on mobile devices, meeting real-time feedback requirements. Memory usage was limited to 325–378 MB.

4.3 User Study

A six-week user study involving 60 tennis enthusiasts demonstrated significant improvements in technical indicators for the experimental group: forehand accuracy improved by 35.8% (control group 19.2%), stroke speed by 27.3% (control group 16.5%), and technical consistency by 41.2% (control group 22.8%). These results confirm the system's effectiveness in accelerating skill improvement.

5. Conclusion and Future Work

This study proposes the Hybrid Perception Attention Network (HPA-Net), an innovative algorithm designed for tennis forehand action recognition and evaluation. By integrating spatial and temporal attention mechanisms with few-shot learning strategies, HPA-Net achieves precise action recognition and evaluation. Experimental results demonstrate its superiority in accuracy, evaluation precision, and computational efficiency. User studies further validate its effectiveness in real-world training.

Despite its strong performance, HPA-Net has several limitations. The model is specialized for forehand analysis and requires significant modifications for other tennis techniques. Performance degrades with lower-quality video inputs (below 720p/30fps), showing a 12.3% accuracy drop with smartphone recordings in poor lighting. The system has limited capability in contextualizing strokes within tactical gameplay, and its few-shot learning approach still depends on carefully selected exemplar demonstrations for initial setup.

Future work will extend the system to other technical actions, explore multimodal perception enhancements, and develop more personalized training plan generation systems.

References

[1] Elliott, B., Reid, M., & Crespo, M. (2009). Technique development in tennis stroke production. ITF Ltd.

[2] Reid, M., Elliott, B., & Crespo, M. (2013). Mechanics and learning practices associated with the tennis forehand: A review. Journal of Sports Science & Medicine, 12(2), 225-231.

[3] Wang, J., Yan, S., Xiong, Y., & Lin, D. (2022). Sports video analysis: Emerging techniques and applications. ACM Computing Surveys, 55(3), 1-34.

[4] Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? A new model and the kinetics dataset. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4724-4732.

[5] Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., & Van Gool, L. (2016). Temporal segment networks: Towards good practices for deep action recognition. Proceedings of the European Conference on Computer Vision, 20-36.

[6] Kovalchik, S., & Reid, M. (2018). A shot taxonomy in the era of tracking data in professional tennis. Journal of Sports Sciences, 36(18), 2096-2104.

[7] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

[8] Zhang, L., Wen, G., & Liu, F. (2023). Deep learning for sports action analysis: Challenges and solutions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(5), 5923-5941.

[9] Laptev, I., Marszalek, M., Schmid, C., & Rozenfeld, B. (2008). Learning realistic human actions from movies. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1-8.

[10] Wang, H., & Schmid, C. (2013). Action recognition with improved trajectories. Proceedings of the IEEE International Conference on Computer Vision, 3551-3558.

[11] Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. Advances in Neural Information Processing Systems, 568-576.

[12] Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local neural networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 7794-7803.

[13] Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). SlowFast networks for video recognition. Proceedings of the IEEE International Conference on Computer Vision, 6202-6211.

[14] Pirsiavash, H., Vondrick, C., & Torralba, A. (2014). Assessing the quality of actions. Proceedings of the European Conference on Computer Vision, 556-571.

[15] Parmar, P., & Morris, B. T. (2019). Learning to score Olympic events. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2449-2458.

[16] Zhu, G., Zhang, L., Shen, P., & Song, J. (2017). Multimodal gesture recognition using 3-D convolution and convolutional LSTM. IEEE Access, 5, 4517-4524.

[17] Vinyals, O., Blundell, C., Lillicrap, T., &

Wierstra, D. (2016). Matching networks for one shot learning. Advances in Neural Information Processing Systems, 3630-3638.

[18] Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. Proceedings of the 34th International Conference on Machine Learning, 1126-1135.

[19] Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L. Y., & Kot, A. C. (2019). NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(10), 2684-2701.

[20] Zhao, Y., Zhang, Z., Wang, Y., & Zhou, J. (2020). Meta-learning for few-shot action recognition. IEEE Transactions on Circuits and Systems for Video Technology, 31(6), 2061-2074.

[21] Faulkner, H., & Dick, A. (2017). TenniSet: A dataset for dense fine-grained event recognition, localisation and description. Proceedings of the British Machine Vision Conference.

[22] Zhou, B., Andonian, A., Oliva, A., & Torralba, A. (2018). Temporal relational reasoning in videos. Proceedings of the European Conference on Computer Vision, 803-818.

[23] Tran, D., Ray, J., Shou, Z., Chang, S. F., & Paluri, M. (2017). ConvNet architecture search for spatiotemporal feature learning. arXiv preprint arXiv:1708.05038.

[24] Yan, S., Xiong, Y., & Lin, D. (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition. Proceedings of the AAAI Conference on Artificial Intelligence, 7444-7452.

[25] Zolfaghari, M., Singh, K., & Brox, T. (2018). ECO: Efficient convolutional network for online video understanding. Proceedings of the European Conference on Computer Vision, 695-712.

