Comparative Analysis of Deep Learning-Based Action Recognition: The example of Table Tennis

Xing Long

College of Intelligent Systems Science and Engineering, Hubei Minzu University, Hubei 445000, China
202112459@hbmzu.edu.cn

Abstract. With the advancement of deep learning technologies, computer vision has shown unprecedented potential in the field of action recognition. Particularly in table tennis, action recognition technologies not only help athletes improve their techniques but also provide real-time feedback during training and competitions. This study thoroughly investigates existing action recognition methods, including dual-stream models, Graph Convolutional Networks (GCNs), and Transformers, and highlights their applications in movement analysis during table tennis activities. The research focuses on revealing the accuracy and real-time capabilities of action recognition to better support coaches and athletes in understanding sports techniques. Additionally, this work introduces that Baidu has developed models capable of recognizing specific table tennis movements, such as serving and returning, with an accuracy rate exceeding 80%, significantly improving the quality and efficiency of training. This paper also discusses the prospective applications of action recognition technology in other sports, as well as potential challenges in future research. It is expected that these advancements will drive the development of sports technology, helping athletes and coaches achieve higher accomplishments through technological means.

Keywords: Action Recognition, Dual-Stream Models, Graph Convolutional Networks, Transformers.

1 Introduction

In recent years, human action recognition has received considerable attention in the field of computer vision. Action recognition is a key technology in computer vision, aimed at using sensors, computer vision techniques, or other related technologies to detect, recognize, and understand human actions and postures. This technology can understand and interpret human actions in video content, and is widely applied in video surveillance, human-computer interaction, virtual reality, and other fields. With the development of deep learning, it has shown significant advantages in image recognition, speech recognition, and natural language processing. In action recognition, deep learning—particularly Convolutional Neural Networks (CNNs),

Y. Wang (ed.), *Proceedings of the 2024 International Conference on Artificial Intelligence and Communication (ICAIC 2024)*, Advances in Intelligent Systems Research 185,
https://doi.org/10.2991/978-94-6463-512-6_29

Recurrent Neural Networks (RNNs), and the recently popular Transformer networks—can effectively process and analyze large volumes of video data, extracting complex temporal and spatial features, thereby improving the accuracy and efficiency of recognition.

Table tennis, a sport that demands high agility and technical skill, poses unique challenges for action recognition. The rapid and small-scale movements of players, frequent and fast interactions between the racket and the ping pong ball, require that action recognition systems accurately capture and analyze movements within a very short time frame. The practical application of table tennis action recognition demonstrates its multifaceted value: first, by analyzing precise action data, coaches can guide athletes to adjust their posture and techniques, optimizing training effects; second, by analyzing competition videos to identify opponents' common tactics and weaknesses, providing tactical advice to athletes; in terms of performance enhancement, athletes can receive immediate feedback through action recognition technology, promptly adjusting their actions to improve performance; finally, the application of action recognition technology has driven the deep integration of traditional sports and technology, enhancing the level of sports competition and the viewing experience.

This paper will delve into table tennis action recognition technology based on deep learning, particularly analyzing and comparing the effectiveness and adaptability of different deep learning architectures in processing table tennis actions.

2 Action Recognition Methods Based on Deep Learning

The research and development of action recognition technology has spanned several decades, evolving from simple pattern matching to complex deep learning models, with increasingly rich methods and applications. Early action recognition systems relied on relatively simple image processing techniques and feature extraction methods such as edge detection and optical flow estimation. These methods, primarily based on manually designed features and shallow learning algorithms, were effective in simple scenarios but often unstable in complex environments. With the improvement of computational power and the increase in data, machine learning, especially deep learning technology, has shown great potential in the field of action recognition. Deep learning models, particularly neural networks, can extract deeper abstract features through self-learning on large datasets, significantly enhancing the accuracy and robustness of action recognition. This ability to automatically learn features from data is unmatched by traditional methods. This article categorizes models into spatial information processing models and temporal information processing models, and describes dual-stream models, graph convolutional networks, and Transformers.

2.1 Dual-Stream Model for Action Recognition

The dual-stream model is a very popular method in the field of action recognition, attracting widespread attention for its effective combination of spatial and temporal information in videos. Early action recognition methods, based on manually designed features and shallow learning models, were somewhat effective in handling simple actions but performed poorly in complex, dynamically changing real-world videos. The purpose of the action recognition method based on the dual-stream model is to more effectively capture the spatial (static) and temporal (dynamic) features of actions in videos [1]. The core idea is to use two parallel CNNs to process video frames, one network processing individual video frames to extract spatial features, and another processing optical flow between adjacent frames to capture dynamic information. This method mainly addresses the problem that a single CNNs model struggles to effectively handle both spatial and temporal information in video action recognition tasks. It includes two main branches: the Spatial Stream Network and the Temporal Stream Network. These two networks can have similar architectures, but they handle different types of input data, focusing on the extraction of spatial features and temporal features, respectively.

First, the video data is split into two streams: the spatial stream and the temporal stream. In the spatial stream, the input is single-frame images, primarily used to capture the spatial information of each frame. In the temporal stream, the input is the optical flow between consecutive frames, focusing on capturing the temporal dynamics of the motion. Both streams have many similarities in structure, undergoing a series of convolutional operations. The advantage of convolutional operations lies in their local connectivity and weight sharing, which help reduce model complexity and effectively extract features while reducing the time required for training. Typically, a normalization layer follows the convolutional layers, aimed at standardizing data to the same scale, which helps accelerate gradient descent and quickly find the optimal solution. After normalization, a pooling layer follows, commonly using max pooling, which extracts the most important features from a specific area, reducing feature dimensions and computational complexity, and helping to prevent model overfitting. After several layers of convolution and pooling, the data flows through two fully connected layers. These fully connected layers flatten and reshape the multidimensional feature vector, where the first fully connected layer primarily enhances the model's generalization capability, and the second adjusts the scale of the output to suit the final output needs. Finally, the output of each stream is classified. The results of these two streams are then fused to produce the final action recognition result.

The dual-stream network is used to analyze the technical movements of athletes in sports such as soccer, basketball, table tennis, and gymnastics, helping coaches and athletes better understand their performance and optimize training. In terms of game tactics, by capturing the positions and movement trajectories of players during matches, coaching teams can understand the execution of game tactics and develop more effective strategies. By recognizing and analyzing athletes' motion patterns, the dual-stream network can identify high-risk movements that lead to injuries, providing

timely feedback to help prevent sports injuries. As a participant in the Media Eval 2022 sports task, a dual-stream network method was proposed for the classification and detection of table tennis strokes [2]. Each stream consists of a series of 3-dimensional (3D) CNNs blocks using attention mechanisms. Each stream processes different 4D inputs. The method utilizes raw image data and posture information, considering the posture as an image by applying it to a black background or to the computed raw Red, Green and Blue (RGB) frames. Optimal performance is achieved by feeding raw RGB data into one stream and posture + RGB (PRGB) information into another stream, and applying late fusion to the features. These methods were evaluated on the provided TTStroke-21 dataset. The report states that the classification accuracy improved to 87.3%, and although the detection results did not surpass the baseline, the Intersection over Union (IoU) was 0.349, and the Mean Average Precision (Map) was 0.110.

A representative work expanded the 2D model into a 3D model by replacing the convolutional kernels of ResNet from two-dimensional to three-dimensional and using 3D pooling, among other changes, introducing a new model architecture [3]. This model can be initialized using 2D pre-trained networks, allowing for deeper architectures without the need for extensive video data for training. It also introduced a new video classification dataset called Kinetics, which is well-balanced in categories, suitably scaled, diverse, precisely annotated, and moderately challenging.

Despite the significant achievements of the dual-stream model in action recognition, it demands considerable computational resources, especially as optical flow computation is typically time-consuming. Moreover, the model requires a large amount of training data to fully leverage the capabilities of deep networks. To address these issues, researchers have proposed various improvements, including real-time optical flow computation methods, a triple-stream model that integrates three different types of data streams (image, optical flow, and skeletal information) to enhance the efficiency and accuracy of action recognition, especially in videos with complex backgrounds [4]. Further integration of deeper temporal information has been explored, attempting to enhance the handling of temporal information in videos by combining CNNs and Long Short-Term Memory (LSTMs) [5]. Another proposed a hidden dual-stream convolutional network, where the temporal stream does not explicitly compute optical flow but instead learns motion information through a sub-network automatically [6]. Previous research introduced a new video representation method that captures spatial and temporal features in videos by integrating Vector of Locally Aggregated Descriptor (VLAD) encoding on top of the dual-stream network [7]. This method significantly improves performance in action recognition tasks and reduces the burden of pre-computing optical flow, enhancing processing speed.

2.2 Graph Convolutional Networks (GCN) for Action Recognition

GCN were developed against the backdrop of handling non-Euclidean data, where in many fields, data points are interconnected through complex relationships and structures. These fields include social networks, knowledge graphs, molecular

structures, and transportation networks, among others. GCN is a type of neural network designed for processing graph data. In the field of action recognition, GCNs are particularly suited for handling complex graph structures composed of human skeletal joints and their connections.

In action recognition based on GCNs, the human body is viewed as a graph, where nodes represent the skeletal joints of the body, and edges represent the physical connections between the joints (such as bones). This graphical representation effectively captures the structural information of the human body, which is particularly important for understanding complex human movements. Graph convolution is the core of GCNs, aiming to extend the traditional convolutional operation (usually applied to regular Euclidean data like images) to handle graph data.

The principle of the GCN is based on the idea of traditional CNNs, expanded to handle graph-structured data. The adjacency matrix A describes the connectivity between nodes in the graph. The feature matrix X, where each row represents a feature vector of a node. GCNs update the feature representation of nodes by applying convolution operations on the nodes. The key to graph convolution lies in how to utilize the local neighborhood information of nodes, aggregating the information of neighboring nodes. The new feature of node i is a function of its own features and the features of its neighbors. This is typically achieved through the product of the adjacency matrix with the feature matrix, namely AX. Then, the aggregated features are transformed through a weight matrix W, which is learned during the training process. These features are usually activated through a nonlinear activation function to enhance the model's expressive power. The entire convolution operation can be represented as:

$$H^{(l+1)} = \sigma(D^{-\frac{1}{2}}AD^{-\frac{1}{2}}H^{(l)}W^{(l)}) \tag{1}$$

, where $A = A + I$ meaning adding the identity matrix I to implement self-connections), D is the degree matrix of A, $\sigma()$ is the nonlinear activation function, $H^{(l)}$ and $W^{(l)}$ are the node representation and weight matrix of the l-th layer, respectively. GCNs can contain multiple convolution layers. Each layer outputs new node features, which provide inputs for the next layer. The output of the last layer can be used for various prediction tasks, such as node-level classification tasks.

Representative works like this proposes a reliability-based feature aggregation mechanism to improve the training efficiency and performance of graph convolutional networks [8]. This method samples neighboring nodes based on the reliability of connections and further aggregates feature information from neighborhoods of different reliabilities through a dual feature aggregation scheme. Additionally, sensitivity to the structure of the graph also has a significant impact, as different graph structures and feature extraction strategies can significantly affect model performance. An adaptive graph convolutional network is proposed, capable of handling graph-structured data of different sizes and connectivity [9]. By learning graph structures effectively through distance metric learning, it demonstrates superior performance on multiple graph-structured datasets. DropEdge technique is proposed, which reduces the overfitting problem by randomly removing edges in the graph

during the training process [10]. This method is similar to the Dropout technique in convolutional neural networks and helps to improve the generalization ability of GCNs on complex graphs.

GCNs are advantageous in handling graph-structured data, effectively capturing the relationships and structural features between nodes in graph data, and are therefore widely used in many application areas. For example, in recommendation systems, users and products can form a bipartite graph, where the products purchased or content of interest by users form the connections in the graph. By learning the associations between users and products, it is possible to predict other products that users might be interested in, thus enabling personalized recommendations. GCNs can also utilize historical behavior data of users and the relationships between contents to identify user interests on platforms like social networks, music, videos, and articles, recommending new content similar to what they have previously browsed. In human pose analysis, the connections between human body joints can be modeled as a graph structure. By learning the interactions between different joints, GCNs can identify and predict complex motion patterns, which are of significant value in sports analysis, surveillance, and health monitoring. The action recognition methods based on GCNs provide a powerful tool for understanding and analyzing complex movements, especially suited for handling complex interactive actions and group activities. The Dynamic Graph Convolutional Neural Network (DGCNN) was proposed for processing point cloud data [11]. Compared to traditional GCNs, DGCNN can capture dynamic information of the topology as it changes over time, making it suitable for tasks such as 3D shape and action recognition, and thus has broad application prospects in fields like sports analysis, surveillance, and health monitoring.

2.3 Transformer

The Transformer is a deep learning model, which is primarily used for sequence-to-sequence tasks such as machine translation and text generation. Before the Transformer, sequence-to-sequence tasks largely relied on Recurrent Neural Networks (RNN). These models performed well in multiple applications, but they had several notable drawbacks, such as sequential dependency—RNNs must process information in sequence order, which leads to inefficient computation during training. They also face problems with vanishing and exploding gradients, which limit the model's ability to learn long-distance dependencies when dealing with long sequences. Additionally, their capacity for parallelization is limited because the processing of sequence data depends on the results of the previous state, making effective parallel processing difficult for RNNs.

The Transformer model was introduced to address these limitations of RNNs and to fully leverage the advantages of the attention mechanism, relying entirely on attention to process sequence data and moving away from the traditional RNN structure. The Bidirectional Encoder Representations from Transformers (BERT) model proposed by [12] significantly improved performance on multiple natural language understanding tasks by pre-training a deep bidirectional Transformer on a large-scale corpus. BERT refined the training methods of the Transformer,

particularly by introducing the concepts of masked language model and next sentence prediction. This brought many advantages; in terms of parallel processing, as the model no longer depends on the previous state of the sequence, different parts of the sequence can be processed in complete parallel, greatly enhancing training efficiency. In capturing long-distance dependencies, the Transformer, through its Multi-head Attention mechanism, can simultaneously focus on multiple positions within the sequence, effectively capturing long-distance relationships. It also offers better flexibility and universality: The design of the Transformer architecture is not only suitable for language processing tasks but can also be extended to many other fields that require processing sequence data. The Reformer is a Transformer model designed to enhance efficiency [13]. It significantly reduces the demand for computational resources by introducing Locally Sensitive Hashing attention and reversible layer techniques, making training on long sequences and large-scale datasets more feasible.

The Transformer introduces a novel method of processing sequence data, primarily relying on the attention mechanism instead of traditional recurrent network structures. The basic architecture of the Transformer model consists of two parts: the Encoder and the Decoder. Each part is composed of multiple identical layers stacked together, each layer including several sub-layers and a residual connection followed by a normalization step. The encoder contains multiple encoding layers, each with two sub-layers, the Self-Attention Mechanism and the Feed-Forward Neural Network. The decoder contains multiple decoding layers, each with three sub-layers: Self-Attention, Encoder-Decoder Attention, and Feed-Forward Neural Network. Self-Attention is the core of the Transformer model, allowing the model to focus on different positions within a sequence while processing it. Self-Attention can be computed in parallel, greatly enhancing the model's efficiency.

The entire process begins with converting each input element of a given input sequence into a fixed-size vector, which is achieved through the Embedding Layer. Then, weight calculations are performed; in the self-attention layer, the input vectors are transformed into three different sets of vectors: Query, Key, and Value. These vectors are obtained by multiplying the input vectors with three trained weight matrices. Attention scores are calculated by computing the dot product between the Query and all Keys, yielding attention scores. These scores determine the degree of focus on other positions at each position.

Due to its efficient parallel processing capabilities and excellent ability to capture long-distance dependencies, the Transformer model has significantly impacted the entire machine learning and artificial intelligence field. This paper introduces the Switch Transformer, a model that supports training with up to trillions of parameters by incorporating sparse activations. The Switch Transformer enhances training efficiency and scalability through model parallelism and routing techniques, becoming the preferred model for processing various sequence data. Models based on the Transformer, such as Google's BERT and OpenAI's Generative Pre-Trained Transformers (GPT) series, have demonstrated powerful capabilities in few-shot learning tasks. For instance, the GPT-3 model presented by [14], which extends the Transformer architecture to a very large scale (175 billion parameters), has achieved unprecedented success in the field of natural language processing, becoming the

cornerstone of many modern Natural language processing (NLP) applications. Furthermore, the concept of the Transformer has also been applied to image processing; the paper applies the Transformer model to image recognition tasks, proposing the Vision Transformer (ViT) [15]. By dividing images into fixed-size blocks ("words"), ViT achieves performance that matches or exceeds the best existing CNN models on image classification tasks.

3 Action Recognition in Table Tennis Activities

3.1 Overview

The application of action recognition technology in table tennis has brought revolutionary changes to training and competition. This section will explore in detail the uniqueness, importance, and application scenarios of table tennis action recognition.

As a sport characterized by extremely fast speeds and high skills, table tennis action recognition possesses several unique features compared to other sports. The actions in table tennis are very fast and brief, and ordinary action recognition systems struggle to accurately capture and analyze these rapidly changing movements. Additionally, the spin and trajectory of the ball in table tennis are complex, demanding higher analytical capabilities from action recognition systems. Moreover, table tennis is highly skillful, with a wide variety of technical actions including various spins, serves, and smashes. This requires action recognition systems to identify and distinguish subtle technical differences.

The integration with action recognition, i.e., table tennis action recognition technology, is very significant for athlete training, competition analysis, and technical improvement: (1) Technical Training: Through action recognition technology, coaches and athletes can receive precise feedback on technical actions, helping them better understand and improve the details of their techniques. (2) Competition Strategy: During competitions, action recognition can be used to analyze an opponent's habits and technical characteristics, thus forming more effective competition strategies. (3) Injury Prevention: Correct action recognition can help athletes avoid sports injuries caused by technical errors.

The application scenarios of table tennis action recognition technology mainly fall into three categories: (1) Athlete Training: Athletes can interact with action recognition systems for technical practice. The system provides real-time feedback to help athletes improve their actions. (2) Online Coaching Platforms: Action recognition technology can be integrated into online table tennis coaching platforms, offering personalized training suggestions and improvement plans to students. (3) Real-time Match Analysis: Analyzing athletes' actions in real-time during matches provides tactical support to coaches.

3.2 Representative Works

In the field of table tennis action recognition, several research and practice projects have demonstrated the potential applications of this technology. For example, the collaborative project between Peking University and Baidu represents a cutting-edge attempt to apply deep learning and video analysis technologies in the sports technology field, particularly in the recognition and scoring of table tennis movements. This project utilizes advanced video processing models—such as the Video Swin Transformer—to enhance the accuracy and efficiency of action recognition.

The Video Swin Transformer is a variant based on the Swin Transformer architecture, specifically optimized for video content. The Swin Transformer itself is a neural network model based on the Transformer architecture, which addresses the issue of long-range dependencies in images by introducing a hierarchical Transformer structure. The Video Swin Transformer extends this model to the video domain, handling spatiotemporal features in video sequences. By expanding the window self-attention mechanism along the temporal dimension, the Video Swin Transformer effectively captures dynamic information within the time series. This enables the model to not only understand the content within a single frame but also comprehend the motion and changes between frames, which is crucial for accurately recognizing the rapid movements in table tennis. Furthermore, the model uses a hierarchical approach to extract features progressively, from local to global, gradually building a comprehensive understanding of the actions. This structure allows the model to maintain efficient computation while handling large-scale video data. The project utilizes the Video Swin Transformer to recognize and classify the movements of table tennis players. By analyzing the athletes' body posture, racket position, and the movement of the ball, the model accurately identifies various types of actions, such as serving, receiving, forehand, and backhand.

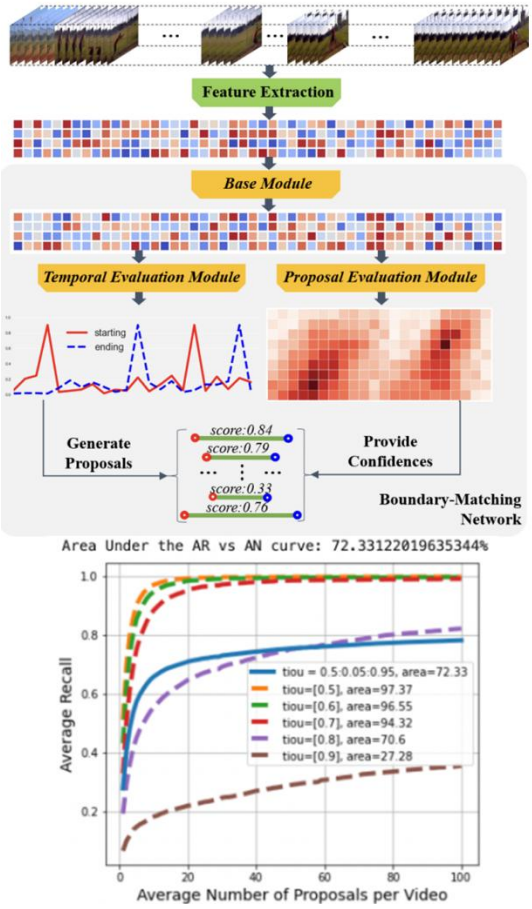

Fig. 1. Baidu's solution for table tennis action recognition [16].

As shown in Fig. 1, based on Baidu's platform for pinpointing table tennis swing actions, compared to mainstream action localization datasets, the best results show ActivityNet1.3 at 67.10% and Thumos14 at 43.54%, with current Area Under Curve (AUC) precision reaching 72.33% on the validation set.

Besides recognizing types of actions, this project also focuses specifically on action quality scoring. This feature is particularly important for athletes' technical training and competition performance analysis. The scoring system is based on the accuracy, fluidity, and adherence to technical standards of the actions. By comparing expert ratings with model outputs, the system has been optimized to ensure the objectivity and accuracy of the scores. Current applications of the project include athlete training, coaching, and competition analysis. Coaches can use this system to provide real-time feedback to athletes, helping them improve their technical movements; competition analysts can use this system for a deeper technical analysis of athletes' performances.

4 Discussion

The rapid development of table tennis action recognition technology has demonstrated significant potential in areas such as training, competition strategy, and sports analysis. However, despite numerous advancements, this technology still faces some limitations and challenges, as well as new opportunities for development.

Advantages: Action recognition technology provides detailed and accurate feedback on athletes' movements, which is very helpful for correcting and improving techniques. Athletes and coaches can use this data to adjust and optimize training plans. Real-time analysis is possible with modern action recognition systems, which can analyze athletes' performance instantaneously during competitions or training, providing real-time feedback. This capability is particularly important for in-game strategic adjustments. The application scenarios for action recognition technology are broadening, from coaching assistance to competition analysis, and even to online teaching platforms, expanding its potential market and application prospects.

Areas for Improvement: There is a strong dependency on data, as the performance of action recognition systems largely relies on a large amount of high-quality training data. The scarcity of data for certain specific or uncommon movements may limit the effectiveness of models. Moreover, the recognition of complex movements still needs improvement; although existing technologies can identify most standard movements, the accuracy and detail in recognizing complex or subtle variations still need to be enhanced. High costs and technical requirements for deploying advanced action recognition systems necessitate expensive equipment and high-level technical support, which may limit their application in low-resource environments.

With rapid technological advancements, several significant improvements are expected in the field of table tennis action recognition over the next few years. Below are some key trends and potential future directions:

Integration of Advanced Machine Learning Models: With progress in artificial intelligence and machine learning, particularly in deep learning technologies, future table tennis action recognition systems will integrate more advanced algorithms and models, such as deeper neural networks and improved Transformer models. These models will better handle high-speed actions and complex interactions, improving the accuracy and response time of action recognition.

Real-time Feedback and Enhanced Training: With improved processing speeds and optimized algorithms, action recognition technology will be able to provide real-time feedback, which is especially important for training and competition. Athletes and coaches will be able to receive instant analysis of action execution, quickly adjusting and optimizing training methods. Additionally, this technology will also support virtual training scenarios, providing athletes with immersive training experiences through virtual reality (VR) or augmented reality (AR) technologies.

Broader Application Scenarios: The application of action recognition technology in table tennis will not be limited to high-level competition and professional training. As the technology becomes more widespread and costs decrease, middle and high schools, amateur sports clubs, and even families might use this technology to enhance table tennis skills and enjoy the sport. Additionally, action recognition can be used for

athlete health monitoring and injury prevention, analyzing athletes' movement patterns and frequencies to promptly detect potential health risks.

Data Privacy and Security: As the application of action recognition technology becomes more widespread, protecting personal privacy and data security becomes a crucial issue. Future systems will need to integrate more advanced data protection measures to ensure all collected and analyzed data complies with privacy standards, and that users have full control over their data.

Interdisciplinary Collaboration: Future table tennis action recognition will be an interdisciplinary field, involving experts in computer science, kinesiology, psychology, and data science. This interdisciplinary collaboration will push the technology to higher levels of development, not just in terms of technical implementation but also in enhancing a deeper understanding of the sport itself.

The future development prospects for table tennis action recognition technology are broad. It will integrate more deeply into all aspects of table tennis, benefiting everyone from amateur enthusiasts to professional athletes. As the technology continues to advance and its applications deepen, table tennis will become more technologically advanced and intelligent.

5 Conclusion

The development and application of table tennis action recognition technology are in a rapid phase of progress. This technology not only provides technological support for professional athletes' training and competition but also offers unprecedented technical assistance to amateur table tennis enthusiasts. Through deep learning and video analysis technologies, such as the Video Swin Transformer, table tennis action recognition can now provide precise action capture, real-time feedback, and action quality scoring, greatly enhancing the efficiency of training and the tactical depth of competitions.

As technology continues to evolve, future table tennis action recognition is expected to become more refined and intelligent, achieving real-time analysis and comprehensive monitoring to help athletes optimize their movements and improve their skill levels. Additionally, as costs decrease and applications become more widespread, more educational institutions and amateur sports organizations will be able to utilize this technology, enabling a broader audience to enjoy the fun and health benefits of table tennis.

Furthermore, facing challenges related to data privacy and security, future action recognition systems need to prioritize the protection of user data to ensure the safety and privacy of the technology. Through interdisciplinary collaboration and technological innovation, the application prospects of table tennis action recognition will become even broader, not only advancing the development of table tennis as a sport but also promoting overall progress in the sports technology field.

References

1. Simonyan, K., & Zisserman, A.: Two-stream convolutional networks for action recognition in videos. Advances in neural information processing systems, 27, 1-9 (2014).
2. Hacker, L., Bartels, F., & Martin, P. E.: Fine-Grained Action Detection with RGB and Pose Information using Two Stream Convolutional Networks. arXiv preprint arXiv:2302.02755 (2023).
3. Carreira, J., & Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299-6308. IEEE, Honolulu (2017).
4. Liang, D., Fan, G., Lin, G., Chen, W., Pan, X., & Zhu, H.: Three-stream convolutional neural network with multi-task and ensemble learning for 3d action recognition. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp. 1-9. IEEE, Long Beach (2019).
5. Ignatov, A.: Real-time human activity recognition from accelerometer data using convolutional neural networks. Applied Soft Computing, **62**, 915-922 (2018).
6. Zhu, Y., Lan, Z., Newsam, S., & Hauptmann, A. Hidden two-stream convolutional networks for action recognition. In Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, pp. 363-378. Springer, Perth (2019).
7. Girdhar, R., Ramanan, D., Gupta, A., Sivic, J., & Russell, B.: Actionvlad: Learning spatio-temporal aggregation for action classification. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 971-980. IEEE, Honolulu (2017).
8. Wang, Y., Li, C., Zhang, J., Ni, P., & Chen, H.: Graph Convolutional Network Using a Reliability-Based Feature Aggregation Mechanism. In Database Systems for Advanced Applications, pp. 536-552. Springer, South Korea (2020).
9. Li, R., Wang, S., Zhu, F., & Huang, J.: Adaptive graph convolutional neural networks. The AAAI Conference on Artificial Intelligence, **32**(1), 1-8 (2018).
10. Rong, Y., Huang, W., Xu, T., & Huang, J.: Dropedge: Towards deep graph convolutional networks on node classification. arXiv preprint arXiv:1907.10903 (2019).
11. Wang, Y., Sun, Y., Liu, Z., Sarma, S. E., Bronstein, M. M., & Solomon, J. M.: Dynamic graph cnn for learning on point clouds. ACM Transactions on Graphics, **38**(5), 1-12 (2019).
12. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018).
13. Kitaev, N., Kaiser, Ł., & Levskaya, A.: Reformer: The efficient transformer. arXiv preprint arXiv:2001.04451 (2020).
14. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D: Language models are few-shot learners. Advances in neural information processing systems, **33**, 1877-1901. (2020).
15. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020).
16. The general action recognition scheme for the four major sports of football, basketb all, figure skating, and table tennis has been made open source. URL: https://ai.baid u.com/support/news?action=detail&id=2744&hmpl=yunying=1.21. Last Accessed 2024/ 05/10.