computers

MDPI

Article

Player's Performance Analysis in Table Tennis Through Human Action Recognition

Kangnan Dong and Wei Qi Yan *

Auckland University of Technology, New Zealand; ksv0677@autuni.ac.nz
* Correspondence: wyan@aut.ac.nz

Abstract: This paper aims to enhance the effectiveness of table tennis coaching and player's performance analysis through human action recognition by using deep learning. In the field of video analysis, human action recognition has emerged as a highly researched area. Beyond post-session analysis, it has the potential for real-time applications, such as providing instant feedback or comparing ideal motions with actual player movements. However, the complexity of human actions presents significant challenges. To address these issues, in this paper, we combine the latest computer vision and deep learning algorithms to accurately identify and classify a few strokes in human action recognition. Throughout in-depth review of the existing methods, we develop a high-precision offline method for player's action recognition. Our experimental results show that the proposed method achieves an average accuracy of 99.85% in recognizing six distinct table tennis actions based on our own dataset.

Keywords: Table tennis; Human action recognition; Deep learning; Computer vision

Citation: To be added by editorial staff during production.

Academic Editor: Firstname Lastname

Received: date
Revised: date
Accepted: date
Published: date

1. Introduction

Human action recognition in sports video analysis has become a critical issue in computer vision and deep learning. This field is crucial for recognizing specific athletic actions, facilitating performance analysis, creating highlight videos, and assisting coaches.

Human action recognition in table tennis has challenges due to the speed of ball and human actions, subtle differences between strokes, the need for precise detection and recognition in quick-moving sports [1-3]. Handcrafted feature-based methods in conventional machine learning have limitations in classifying human actions in sports. This foundation led to further innovations, including the Inflated 3D ConvNet (I3D) in 2017, which extended 2D CNN architectures along the temporal dimension and proved effective on large-scale action datasets like Kinetics [4]. To address these challenges, we propose the methods by using the state-of-the-art methods in computer vision and deep learning, focusing on Transformer models to accurately recognize human actions in table tennis games [5].

This research project aims to develop a method capable of identifying specific strokes in table tennis. Consecutive video frames of players' actions are analyzed to ensure accurate classification, providing efficient post-session feedback for coaching purposes.

Another key aspect of this method is the use of Google MediaPipe platform for pose estimation from pre-recorded videos [6]. We utilize the platform for human pose estimation due to its ability to precisely detect key joints of a human body, which are essential for accurately identifying player's actions in table tennis games. By tracking these key points across multiple frames, our methods are able to recognize subtle variations between player's actions, such as forehand drive and smash.

The players and coaches in table tennis games need detailed insights into performance after games, which requires software that can accurately classify human actions and efficiently handle recorded videos. The proposed method copes with these challenges by integrating Transformer models and MediaPipe platform, especially for human pose estimation, delivering precise action recognition from recorded videos, enabling coaches to provide detailed feedback after the matches [7].

Additionally, using MediaPipe for pose estimation provides valuable feedback. By accurately detecting the key joints, it assists us in analyzing player's actions, offering insights into specific performance. The platform ensures smooth and accurate analysis of sport videos, further enhancing practical applications.

This research work contributes to the field of human action recognition in table tennis by utilizing a Transformer-based approach combined with pose estimation, the accurate and efficient method for human action recognition is offered for timely feedback after the games. The adaptability and accuracy make it suitable not only for table tennis but also for other quick-moving sports that require detailed motion analysis.

In this paper, we will introduce the related work in Section 2. Our methods are depicted in Section 3. The experimental results are demonstrated in Section 4. The conclusion will be drawn in Section 5.

2. Related Work

Human action recognition in sport games has gained significant attention, particularly in quick-moving sports like table tennis, where fine-grained actions are difficult to be captured. Deep learning, especially with Transformer architectures, has led to significant advancements in recognizing and classifying human actions with better accuracy, compared to conventional methods in machine learning with handcrafted features [8].

Conventional machine methods for human action recognition in sports relied heavily on handcrafted features [9]. These approaches often employed motion and appearance descriptors, such as space-time interest points (STIPs) and dense trajectories [10]. While being effective, these methods were less adaptive and reliable in sport games where human actions are rapid, subtle, and often similar. In response to these constraints, early deep learning models like 3D convolutional neural networks were developed to capture both spatial and temporal features [11]. Table tennis, with player's dynamic movements, shows a particular challenge due to the fine distinctions between different actions.

Previous studies have utilized MobileNetV2 for efficient feature extraction and Transformer models for temporal modeling. Building upon this, we propose a dual-output model that combines these methods to achieve both accurate action classification and boundary detection [12].

MobileNetV2 provides efficient feature extraction with low computational costs [13]. By reducing the complexity of convolutional layers, MobileNetV2 significantly diminishes computational costs and memory demands, making it particularly suitable for handling large volumes of video frames from labelled datasets. Accurate feature extraction from each frame is crucial in human action recognition, the visual features extracted by using MobileNetV2 ensure that subtle actions in table tennis games, such as racket rotation and player posture changes, are captured and identified effectively.

Transformer processes the feature sequence through its self-attention mechanism, capturing both short-term and long-term dependencies. The proposed dual-output model processes the feature sequence through its self-attention mechanism, capturing both short-term and long-term dependencies. This model includes one branch for human action classification and another for action segmentation, i.e., boundary detection. By combining these outputs, the model achieves accurate human action classification while also identifying action boundaries in sports videos.

Transformer models broke down input video frames as a series of patches, turning them into vectors, and treating them like tokens. Transformer models, like Vision Transformer (ViT) [14] and TimeSformer [15] have significantly improved human action recognition tasks by effectively leveraging both spatial and temporal features [1]. These approaches are particularly effective in sports involving dynamic human actions.

Swin Transformer has achieved a Top-1 accuracy exceeding 84.9% on general datasets like Kinetics-400 [16]. but its performance is constrained in domain-specific tasks like table tennis due to the challenges of fast-paced movements and subtle variations. In table tennis, the difficulty lies not only in the speed of the actions but also in the subtle variations. For instance, distinguishing between a forehand drive and a smash requires precise attention to the player's body movements and racket trajectory. These intricacies demand models capable of fine-grained action recognition.

Our proposed method addresses these challenges by achieving human action recognition accuracy 96% for table tennis games, significantly outperforming general Transformer-based models. The model combines accurate classification of human actions with action boundary detection, providing comprehensive post-action analysis based on labeled video data. Furthermore, the algorithm processes frames at an average speed 18.3 milliseconds per frame based on an NVIDIA RTX 3070 GPU. This balance between high accuracy and computational efficiency makes the model well-suited for offline coaching applications, where detailed feedback on player actions is invaluable for improving training effectiveness and game strategies. Transformer-based parallel processing capabilities ensure scalability for analyzing large volumes of offline video data [17].

Combining deep learning with human body tracking is crucial for improving sports performance analysis. Transformer models, capable of classifying spatial and temporal patterns, represent a significant advancement in recognizing actions in dynamic sports like table tennis.

3. Methodology

The focus of our research project is on designing and developing a method for human action recognition in table tennis—a sport game known for its rapid, precise, and often subtle strokes. Inspired by the speed and intricacies of table tennis, we create a deep learning model that could keep pace with the players while accurately distinguishing between various actions. By harnessing cutting-edge deep learning models, our method was designed to not only detect but also classify the fast motions.

To enhance the effectiveness of our approach, we integrated a Transformer-based deep net for human action recognition, which ensures that our methods can respond quickly to the fast-moving actions while accommodating subtle variations of those actions that are characteristic of table tennis games.

3.1. Data Collection and Augmentations

To acquire visual data for this project, we recorded videos of six specific actions performed by players in table tennis games, we supplemented these recordings with online training videos. This approach allows us to capture human actions across various environments, while increasing the model adaptability and robustness. We collaborated with professional table tennis coaches to record our videos by using a handhold camera operating at 30 frames per second (fps) from an umpire’s angle of view. This setup ensured that subtle motions of players and path of ping-pong ball were accurately recorded.

Our dataset includes six types of human actions performed by two players—a coach and the author: Backhand Drive, Forehand Drive, High Toss Loop, Long Push, Short Placement, and Smash. Additionally, a "NoAction" class was added to represent moments where no specific action was performed. This class includes preparatory actions and other frames not directly related to the main actions (starting, hitting, and end frames). Each action is annotated to ensure comprehensive coverage of key frames within each action.

Regarding visual feature extraction, we employed MobileNetV2, a lightweight architecture balancing accuracy and computational efficiency. MobileNetV2 excels in extracting spatial features from video frames, reducing computational costs while maintaining high accuracy. The inverted residual structure and linear bottleneck layers of the model enable efficient processing of high-speed human motions, making it highly effective for analyzing dynamic actions in table tennis. This lightweight design minimizes computational complexity, ensuring accurate player performance analysis while maintaining fast processing speeds—an essential requirement for providing timely feedback during coaching sessions.

We utilized OpenCV platform to extract video frames from the recorded videos. Each frame was resized to 224×224 pixels to match the input size required by MobileNetV2, converted from BGR to RGB, and normalized it to ensure consistency in model input. To enhance the robustness and mitigate overfitting, particularly for underrepresented classes, we applied data augmentation methods such as horizontal flipping and slight rotation. These pre-processing steps enabled the model to detect player's styles across different environments while also evaluating its processing capabilities, which are critical for coaching applications.

All data was annotated with professional players and coaches in table tennis to ensure that the starting, hitting and end frames were accurately labeled. We employed a two-stage review process, with initial annotations conducted by well-trained annotators and final reviews completed by professional coaches to ensure accuracy and consistency.

While maintaining a balanced dataset, we collect approximately equal number of samples for each class of strokes. However, due to dynamic nature of table tennis, a few classes of human actions naturally occurred frequently.

To ensure a balanced dataset, we recorded videos for each stroke class, aiming to collect a comparable amount of footages for six actions: Backhand Drive, Forehand Drive, High Toss Loop, Long Push, Short Placement, and Smash. Additionally, a "NoAction" class was included to represent moments without specific actions, such as preparation, starting, hitting, and ending phases.

All videos were recorded at a consistent frame rate of 30 frames per second, resulting in a total of approximately 36,000 frames. Each action lasted approximately 15 to 23 seconds, depending on the speed and complexity of the movement, with a total video duration of around 20 minutes. The dataset was divided into training, validation, and testing sets, with 70% frames allocated for training, and 15% for validation and testing. The validation and test sets included videos performed by players excluded from the training set, ensuring independence in evaluation.

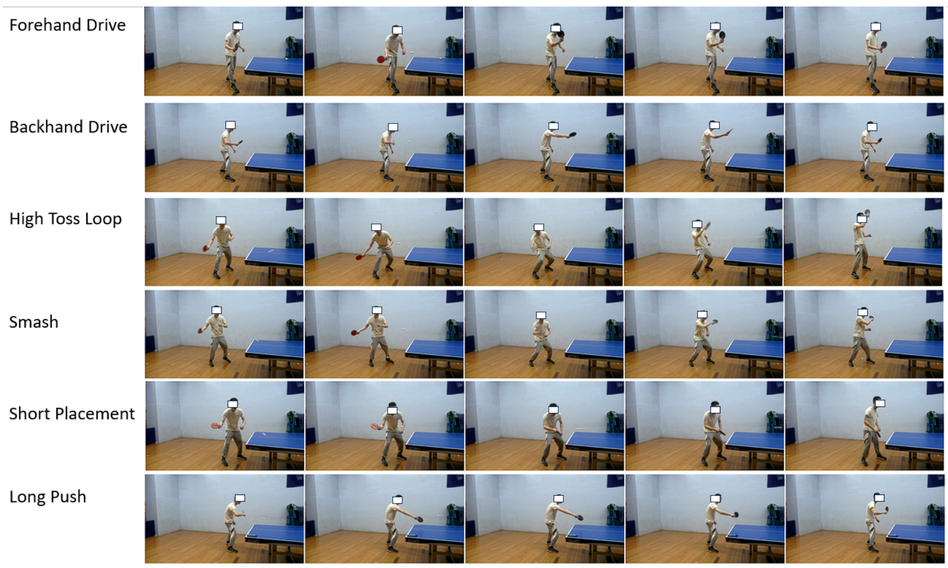

Figure 1. The samples from our training dataset, showing 5 consecutive frames for each of the six player's actions.

3.2. Pose Estimation Using MediaPipe

Estimating poses is an integral component of human action recognition in table tennis, which is a necessary step for classifying player's action and recognizing patterns of human actions. In this paper, MediaPipe platform was employed owing to its high accuracy, fast performance, and ultra independence.

As determined by biomechanical studies of table tennis, further processing covers the detection of key points of player's wrist, elbow, shoulder joints, hip joints, and knee; foot ankle and head position. These key points are integral for accurately representing human actions of table tennis players.

Specific key points are recorded across consecutive frames to capture the temporal characteristics of human actions, balancing computational costs with accuracy, provided key data about the temporal patterns that allows us to distinguish between human actions that might appear similar spatially but differ considerably in the temporal.

3.3 Network Architecture

Our proposed model for human action recognition was created to accommodate both spatial and temporal dependencies in a sequence of video frames. Our architecture makes use of MobileNetV2 for visual feature extraction while Transformer-based models handle temporal sequence, this combination has enabled us to detect 6 distinct player's actions simultaneously while simultaneously segmenting the actions.

Video data is firstly extracted by using OpenCV platform, followed by visual feature extraction via MobileNetV2. The extracted features were organized into a sequence to capture the temporal dependencies essentially for accurately recognizing player's actions in table tennis. Transformer-based models tackle the sequence for accurate human action segmentation and recognition, with built-in counting function ensuring that each instance of live play is accurately counted. For each frame t in the video sequence, MobileNetV2 extracts a 1280-dimensional feature vector as shown in Eq.(1).

$$F_t = MobileNetV2(frame_t). \tag{1}$$

These feature vectors capture the spatial information of the video frames. To model the temporal dynamics inherent in consecutive frames, the extracted feature vectors are organized into a sequence, as defined in eq.(2).

$$S = [F_{t-n}, F_{t-(n-1)}, \dots, F_t]. \tag{2}$$

Since Transformers do not inherently understand the order of input frames, we adopt positional encodings to represent the sequential order. This is completed by adding positional encoding vectors to each feature vector F_t, where each position has a unique encoding method based on sine and cosine functions of different frequencies. This allows us to capture the temporal sequence as shown in Eq.(3).

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d}}\right), \quad PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{\frac{2i}{d}}}\right). \tag{3}$$

where *pos* represents the position in the sequence, d is the dimension of positional encoding.

We chose Transformer models due to the ability to efficiently capture long-range dependencies through parallel processing, which is crucial for recognizing player's actions. Moreover, Transformers avoid the vanishing gradient problem, making them more suitable for capturing the fast and subtle actions.

In addition, the use of consecutive frames instead of a fixed number of frames allows the model to be flexibly adaptive to varying lengths of sequences, enhancing its ability to generalize across different contexts and action classes. This flexibility is especially important while classifying player's strokes that may have different execution times, making the model more robust in handling diverse inputs.

The sequence of feature vectors, now with positional encodings, is processed by using Transformer. The core component is the Multi-Head Self-Attention mechanism, which calculates the attention score for each frame. The self-attention is defined as eq. (4).

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V. \quad (4)$$

where Q (queries), K (keys), and V (values) represent projections of the input sequence S, $d_k$is the dimension of keys. The Multi-Head Attention mechanism allows the model to focus on multiple parts of sequence at once, capturing both short-term and long-term dependencies much effectively. Transformer has 8 attention heads, each with a size of 64 dimensions. After the self-attention mechanism, the resulting attention scores are passed through a position-wise feedforward network, consisting of two fully connected layers with ReLU activation. The feedforward network is applied to each position independently.

$$FFN(x) = \max(0, xW_1 + b_1)W_2 + b_2. \quad (5)$$

Residual connections are added around both the self-attention and feedforward layers to stabilize the training process. These residual connections make the network much robust. Each sublayer is followed by layer normalization.

Table 1. Architecture of the proposed model

Layers	Output shapes	Param #
InputLayer	(None, 8, 1280)	0
MultiHeadAttention	(None, 8, 1280)	2,624,256
LayerNormalization	(None, 8, 1280)	2,560
Dropout	(None, 8, 1280)	0
Dense	(None, 8, 128)	163,968
LayerNormalization	(None, 8, 1280)	2,560
Dropout	(None, 8, 1280)	0
GlobalAveragePooling1D	(None, 1280)	0
Dense	(None, 64)	81,984
class_output (Dense)	(None, 13)	845
flag_output (Dense)	(None, 3)	195

Table 2. Summary of training settings

Training Information	Values
Total Params	3,123,472 (11.92 MB)
Trainable Params	3,123,472 (11.92 MB)
Epoch	1/300
Step Time	3s 42ms/step
Class Output Accuracy	0.09%

In Table 1 and Table 2, the architecture of our proposed model is illustrated, showcasing how the model copes with the features from consecutive frames to produce outputs. The model was trained by using the Adam optimizer with sparse categorical cross entropy as the loss function, which is particularly suitable for multiclass classification tasks with integer encoded labels. The training process was conducted over 300 epochs with a batch size of 32. The model has approximately 3.1 million parameters, ensuring it can deal with the complex movements in table tennis. The key components include:

- Multi-Head Attention: With 8 Heads and 64 dimensions per Head, allowing the model to focus on temporal parts of the input sequence.
- Fully Connected (Dense) Layers: These layers take use of the ReLU activation function to introduce non-linearity, enabling the model to learn complex patterns, while the dropout is applied to reduce overfitting and improve generalization.

4. Experimental Results

The Transformer-based model for human action recognition in table tennis games was evaluated by using a comprehensive set of metrics. In this section, we present the overall accuracy, training, and validation progress, as well as per-class performance of the proposed model. Figure 2 illustrates the learning curves for both human action classification and action boundary detection over 300 epochs.

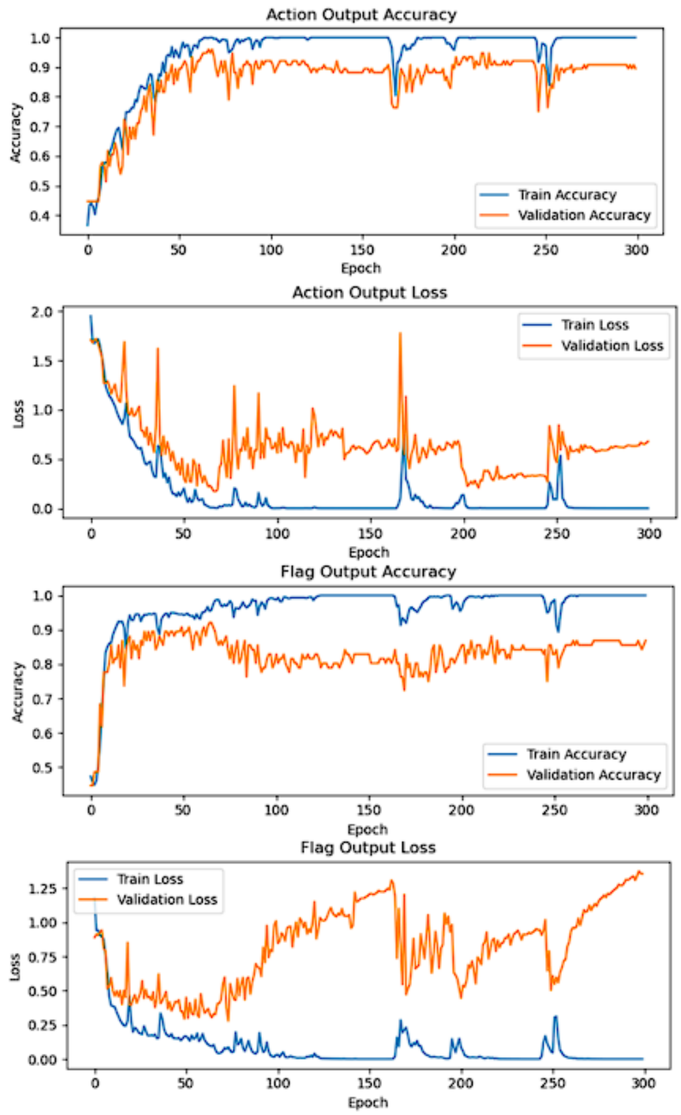

Figure 2. Training / validation accuracy and loss for human action classification as well as action boundary detection.

The model demonstrated consistent improvements during the training process. The training accuracy steadily increased, while the training loss showed a gradual decrease, indicating effective learning of spatial and temporal features from the data. Despite periodic fluctuations in validation loss after epoch 150, the overall trend stabilized toward the later stages of training, suggesting that the model successfully generalized to unseen data. These fluctuations are likely due to the inherent complexity of certain stroke classes, such as visually similar actions like Forehand Drive and Short Placement, which challenge the model's ability to distinguish fine-grained temporal features.

The model achieved human action classification accuracy 96%, highlighting its ability to differentiate between various strokes such as Forehand, Backhand, and Smash. These results underscore the robustness of the Transformer architecture in capturing both the spatial and temporal dependencies of human actions

The macro-average F1-score 0.93 reflects the balanced performance across all classes, demonstrating its ability to classify less frequent actions such as "HighTossLoop" (F1-score: 0.91) and "ShortPlacement" (F1-score: 0.86) with notable accuracy. In contrast, the weighted-average F1-score 0.96 highlights the strong performance on frequent actions, particularly "NoAction", which achieved F1-score 0.99. These results underline the effectiveness of this proposed model in addressing the challenges posed by considering class imbalances, a common issue in real-world datasets.

The action output loss curve reveals steady progress during training, with the loss decreasing consistently as the epochs advance. While validation loss exhibited occasional spikes after epoch 150, it ultimately stabilized toward the end of training. This fluctuation likely stems from the complexity of distinguishing visually similar strokes, such as Forehand Drive and Short Placement. Despite these variations, the model demonstrated strong generalization capabilities without significant overfitting.

With an overall classification accuracy 96%, the model effectively captured both spatial and temporal dependencies in table tennis actions. However, transitional phases, including starting and hitting frames, presented challenges, reflected in lower F1-scores for these segments. Addressing these challenges through improved data representation or augmentation strategies could further enhance performance across all action phases.

Overall, the Transformer-based model has proven highly effective in human action classification, with potential for further refinement in action boundary detection and action transition detection. The enhanced training strategies, such as augmenting the data to better capture middle phases or incorporating context-aware features, could address the fluctuations seen in the validation loss and improve the model performance.

The self-attention mechanism in the Transformer enabled it to focus on relevant parts of a sequence, reducing misclassification and improving accuracy for complex action. This led to higher precision and recall for advanced actions, ultimately enhancing overall performance.

In Figure 3, the confusion matrix for player's action classification is presented. The model performs well across most of given actions, with high precision and recall, though misclassifications remain uncertainty.

The performance of the proposed model was evaluated based on the test set, and the detailed metrics were computed to assess how well the model distinguishes between different table tennis actions. In Figure 4, the overall accuracy for human action classification was 96%. While the model performed very well across most classes, lower recall was observed for human actions like Long Push, where the recall was 0.75, indicating that the model struggles to correctly identify all instances of this action.

Pertaining to human action segmentation, the model achieved an accuracy 87%, with strong performance in detecting the starting, hitting and end time of each action. However, the detection of the middle phase showed lower recall, indicating the room for improvement in distinguishing this transitional phase.

The confusion matrices and classification provide deeper insights into the model strengths and areas for improvement. While the model excels in distinguishing between most actions, further refinement may be required to improve its performance in differentiating Forehand Drive from NoAction and resolving the confusions observed in Long Push.

Table 3. Performance metrics for latency and standard deviation.

Action Types	Average Latency (ms)	Standard Deviation (ms)
Short Strokes	157	23
Medium Stokes	213	31
Long Stokes	286	42
Serves	198	28

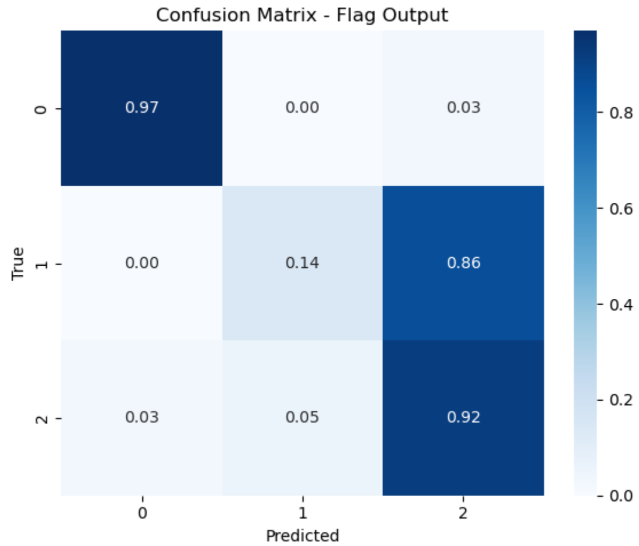

Figure 3. Confusion matrix for action segmentation.

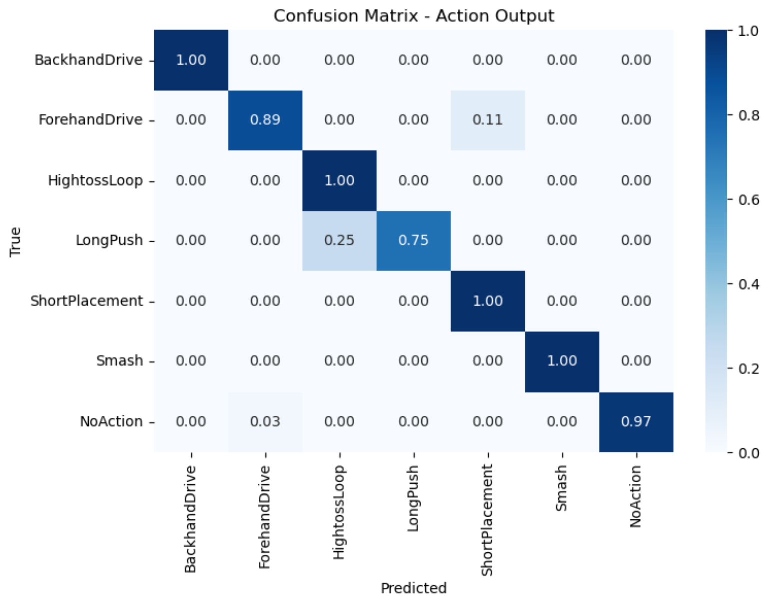

Figure 4. Confusion matrix for human action classification.

Table 4. Performance metrics of average accuracy.

Metrics	Values

Average Processing Time per Frame	18.3 ms
Action Recognition Accuracy	91.2%
Maximum Consecutive Frames Processed	3600
System Stability Duration	120 minutes

Temporal performance is crucial for practical applications in coaching and player analysis. This section presents a comprehensive analysis of the temporal characteristics of the proposed model.

In Table 4, the results show that the latency varies across different classes and durations of human actions. Latency in this study is defined as the time taken by the model to process a sequence of input frames and produce a classification output, focusing solely on inference time. For simpler actions such as Flicks and Flips, the average latency was 157ms, reflecting shorter temporal dependencies and lower computational demand. In contrast, more complex actions such as Loops and Smashes exhibited a higher average latency of 286ms, due to their richer temporal patterns requiring the analysis of longer sequences for accurate classification.

The latency values demonstrate the model's efficiency in handling sequential data for offline action recognition, as tested on 3,600 frames corresponding to a 2-minute video recorded at 30 fps. While effective for post-session evaluation, optimizing latency could make the system adaptable for real-time applications, such as providing immediate feedback during coaching sessions. The results highlight the model's adaptability to varying input sequences, ensuring robustness across diverse scenarios.

Table 5. The results of our developed method.

Frame Rate (fps)	Recognition Accuracy (%)	CPU Utilization (%)	GPU Utilization (%)
15	88.7	23	31
30	91.2	37	58
60	93.5	63	82
120	94.1	89	95

Table 5 shows the performance at various frame rates, illustrating the trade-offs between recognition accuracy and computational workload. The recognition accuracy increased from 88.7% at 15 fps to 94.1% at 120 fps, likely due to better temporal details captured with higher frame rates. However, 30 fps was found to provide an optimal balance between processing efficiency and recognition accuracy for offline analysis.

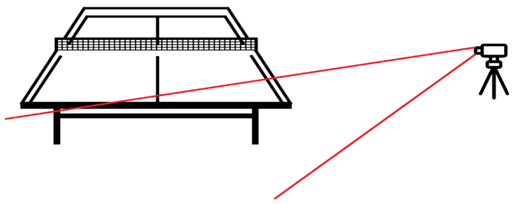

Figure 5. The view angle of a camera to capture the player’s actions.

Figure 5 shows the camera setup to capture the player's actions during data collection time. The camera is positioned at a height above the table, approximately 2 meters away, angled at 45 degrees to capture the entire body of the player. This setup ensures an optimal view angle of the player’s movements and ball trajectory, providing comprehensive data for subsequent action analysis. The paddle faces the camera, allowing for a clearer observation of the strokes, which helps in accurately analyzing the player's performance through MediaPipe platform.

Table 6 presents the average statistics of six classes of player’s actions and a "NoAction" class, indicating the ability to correctly identify each action. For instance, the action Forehand Drive consistently achieves an average probability above 99%, which highlights the precision and reliability of the model in detecting this action without missing any key movements.

Table 6. The probability for each class of human actions.

Actions	Average Statistics
Backhand Drive	99.90%
Forehand Drive	99.92%
High Toss Loop	99.85%
Long Push	99.88%
Short Placement	99.89%
Smash	99.91%
No Action	99.87%

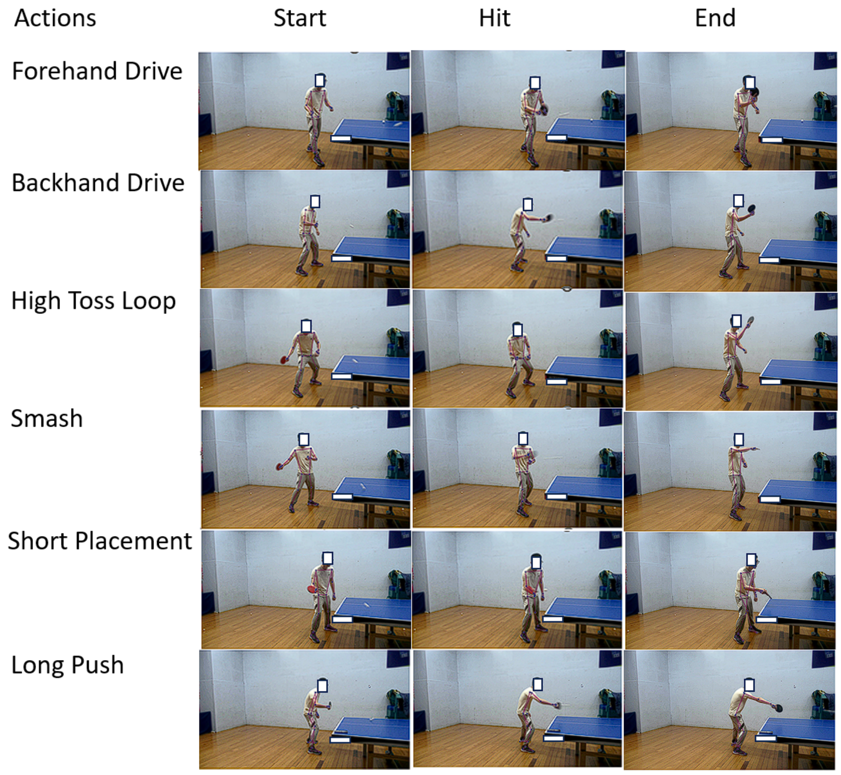

Figure 6. The detection of all human actions.

The proposed model was tested across various human actions in table tennis games, including the actions: Smash, High Toss Loop, Long Push, Backhand Drive, Forehand Drive, and Short Placement as shown in Figure 6, which illustrates the six human actions in three distinct phases—Starting, Hitting, and End time. Each stroke is represented by using a sequence of video frames.

The consistent detection across different stroke classes, including class "NoAction", highlights the robustness and adaptability of our proposed method. It successfully managed variations in player’s actions, lighting conditions, and stroke speeds without compromising accuracy. The timely feedback provided by the proposed method simplifies coaching and training, enabling immediate performance review and adjustments.

5. Conclusions

This paper has demonstrated the potentiality of advanced deep learning methods, particularly Transformer models, for enhancing player's action recognition in table tennis. The proposed method provides a robust solution for human action recognition, offering comprehensive feedback for players and coaches after reviewing the recorded training sessions.

The integration of Transformer models and MobileNetV2, with the ability to capture both spatial and temporal dependencies, has proven effectiveness in accurately classifying various strokes in table tennis. The proposed method ensures that it can be applied to practical training process, where efficient and precise feedback based on post-session video analysis is critical. Additionally, the pose estimation enhances the accuracy of the proposed model by tracking key points of human body, further improving human action recognition.

Our future work should focus on expanding datasets and incorporating more robust methods to handle variations in lighting conditions, camera view angles, and player's movements [18]. Despite these challenges, this research paper provides a solid foundation for human action recognition in table tennis games and has the potential to be adapted for broader applications beyond table tennis.

References

1. Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; Schmid, C. ViViT: A video vision transformer. In IEEE/CVF International Conference on Computer Vision, 2021, pp. 6836–6846.
2. Bertasius, G.; Wang, H.; Torresani, L. Is space-time attention all you need for video understanding? In International Conference on Machine Learning, 2021, pp. 813–824.
3. Summerfield, M. *Rapid GUI Programming with Python and Qt: The Definitive Guide to PyQt Programming*; Pearson Education, 2015.
4. 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, 2017, pp. 6299–6308.
5. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Houlsby, N. An image is worth 16×16 words: Transformers for image recognition at scale. In International Conference on Learning Representations, 2021.
6. Lugaresi, C.; Tang, J.; Nash, H.; McClanahan, C.; Uboweja, E.; Hays, M.; Grundmann, M. MediaPipe: A framework for building perception pipelines. *arXiv* 2019, arXiv:1906.08172.
7. Kulkarni, K.M.; Shenoy, S. Table tennis stroke recognition using two-dimensional human pose estimation. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 4576–4584.
8. Zhou, H. Computational Analysis of Table Tennis Games from Real-Time Videos Using Deep Learning. Master's Thesis, Auckland University of Technology, New Zealand, 2023.
9. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3D convolutional networks. In IEEE International Conference on Computer Vision, 2015, pp. 4489–4497.
10. Wang, H.; Kläser, A.; Schmid, C.; Liu, C. L. Dense trajectories and motion boundary descriptors for action recognition. International Journal of Computer Vision, 2016, 103(1), 60–79.
11. Ji, S.; Xu, W.; Yang, M.; Yu, K. 3D convolutional neural networks for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(1), 221–231.
12. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4510–4520.
13. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
14. Touvron, H.; Cord, M.; Dei, T.; Matthey, L.; El-Nouby, A.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training data-efficient image transformers & distillation through attention. In Proceedings of the International Conference on Machine Learning, 2021, pp. 10347–10357.
15. Zhang, C.; Wei, X.; Wang, Y.; Jin, R.; Yan, S. TimeSformer: Temporally-coupled Transformer for video action recognition. In IEEE/CVF International Conference on Computer Vision, 2021.
16. Liu, Z.; Ning, J.; Cao, Y.; Wei, Y.; Zhang, Z.; Lin, S.; Hu, H. Video Swin Transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 3202–3211.
17. Girdhar, R.; Ramanan, D. ActionVLAD: Learning Spatio-temporal Aggregations for Action Classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017, pp. 971–980.
18. Yan, W. Computational Methods for Deep Learning: Theory, Algorithms, and Implementations, 2023, Springer.