Automatic Segmentation of Table Tennis Match Video Clips Based on Player Actions for Enhanced Data Acquisition

Zhong-Kai Wei and Jieh-Ren Chang *

Department of Electronic Engineering, National Ilan University, Ilan County, Taiwan
Email: weicary28@gmail.com (Z.-K.W.); jrchang000@gmail.com (J.-R.C.)
*Corresponding author

***Abstract*—The collection of data in table tennis competitions is crucial for professional coaches and top players. Big data provides them with insights to analyze opponents' strategies and adjust their own training directions. However, existing models for recognizing table tennis serving and receiving actions still suffer from low accuracy, making them unsuitable for automated match data recording. Currently, training data for machine learning models are sourced from manually edited videos with labels, which is time-consuming, requires professionals, and generates limited data, leading to poor accuracy in existing action recognition models. Therefore, this study aims to achieve automatic video clipping of table tennis matches using computer vision and machine learning techniques to enhance data collection efficiency. This innovative approach reduces the need for human resources, speeds up data generation, and ensures high accuracy in segment identification. This research effectively automates the labor-intensive process of data preparation, solving the problem of obtaining large, annotated datasets for training machine learning models. Furthermore, this automated clipping framework can be applied to other action recognition tasks, improving the accuracy of semi-supervised learning models in various domains. The proposed Table Tennis Playing Status Recognition Model (TTPSRM) achieved a serve recognition accuracy of 98.4% and a receive recognition accuracy of 96.6%. The automated video segmentation system, tested with 2022 International Table Tennis Federation (ITTF) competition videos, successfully clipped 70% of serve actions, with average start and end point errors of 0.547 and 0.287 s, respectively.**

***Keywords*—table tennis, Gated Recurrent Unit (GRU), human pose estimation, OpenPose, automatic clipping**

I. Introduction

In recent years, machine learning technology has seen widespread application across various sports disciplines, enhancing the analysis and training of athletes [1–3]. In sports such as baseball and basketball, machine learning aids in analyzing players' techniques and strategies, thereby improving performance [4, 5]. However, in table tennis, the application of deep learning techniques for players' action information collection has been limited due to the reliance on manual data collection [6]. Professionals currently watch broadcast videos to manually record serve and receive actions, ball landing points, and scoring information. This manual approach is time-consuming, labor-intensive, and prone to subjective biases, making it challenging to conduct effective big data analysis.

Action recognition is pivotal for automating match information recording. Yet, existing action recognition models lack the accuracy in match broadcasts due to insufficient training data [6]. Deep learning models depend heavily on large volumes of accurately labeled training data. The shortage of data results in poor model performance in identifying players' actions [7]. The goal of this research is to develop an automated system to extract players' actions clip from complete table tennis match videos. This system aims to generate a large amount of training data to improve the accuracy of action recognition models.

This research purposes to enhance the efficiency and accuracy of data collection in table tennis by utilizing computer vision and machine learning technologies. Specifically, it focuses on automatically extracting serve and receive action segments from match videos to address the need for large, high-quality training datasets. The proposed system will employ a Table Tennis Playing Status Recognition Model (TTPSRM) that uses a human pose estimation model, OpenPose, combined with a Gated Recurrent Unit (GRU) deep learning network to identify players' action statuses. It will then pass through a continuous video action recognition system and an action clip automatic cutting system. These systems will recognize and clip serve and receive actions from the videos. Thus, it facilitates the creation of extensive training datasets necessary for deep learning.

By improving the efficiency of training data preparation, this research seeks to overcome the limitations of current manual methods, which are inefficient and reliant on professional expertise. The automated system will enable faster and more accurate data collection, thereby

Manuscript received July 10, 2024; revised August 10, 2024; accepted September 9, 2024; published December 12, 2024.

supporting more advanced and effective research and development in table tennis.

II. Related Work

A. Action Classification

According to current research, studies available on GitHub have utilized OpenPose in combination with Support Vector Machine (SVM), Convolutional Neural Network (CNN), and Long Short-Term Memory (LSTM) [8]. These studies extract skeletal information of players from video and use classifiers to categorize actions. This approach accurately identifies whether a player's posture in the frame indicates a forehand or backhand, enabling proportional statistical analysis.

In baseball, recent research also employs OpenPose technology to collect skeletal information of batters' swings. This method captures the limb movement behaviors and angle information of the batter, enabling an understanding of the relationship between the swing posture and the flight distance of the ball after being hit [5].

The aforementioned studies indicate that current research on action recognition predominantly focuses on dynamic sports categories. However, there is a lack of automated action recognition models specifically for table tennis. Therefore, the objective of this research is to develop a machine learning model for the action recognition of table tennis players and to automate the segmentation of match video. This aims to enhance the effectiveness of machine learning research in the domain of table tennis.

B. Gated Recurrent Unit (GRU)

The Gated Recurrent Unit (GRU) is a type of Recurrent Neural Network (RNN) particularly well-suited for handling sequential data [9, 10]. Compared to traditional RNNs, GRUs have a simpler architecture and more efficient computation capabilities, exhibiting superior performance in processing long-term sequential data [11]. GRU neural networks consist of two gates: the update gate and the reset gate. These gates control the flow of information, addressing the vanishing gradient problem commonly associated with traditional RNNs. Therefore, it can maintain robust performance over extended sequences. The update gate determines the extent to which the previous time step's state influences the current state. The reset gate controls the relevance of the previous time step's state in generating the hidden state [12].

In action recognition, GRUs can effectively process the skeleton point information generated by OpenPose [13–16]. These skeleton points are coordinate sequences that describe human postures and motion changes. By using these skeleton points as inputs to the GRU, it processes video sequences from various datasets. The GRU model can accurately recognize a range of complex action patterns, such as running, jumping, and waving, enabling the model to learn the continuous movement patterns of skeletal coordinates for action classification.

III. Research Method

A. Table Tennis Serve and Receive Action Database

The training data used in this study is from a self-established table tennis serve and receive action database, named Table Tennis Action 20 (TTA 20). The TTA 20 database was created to support the training of models for recognizing serve and receive actions. In collaboration with table tennis professionals, the data was collected from online platforms featuring broadcast of matches played by the top ten ranked male players in the world. These full match videos were manually clipped to extract segments of serve and receive actions. Each segment was then labeled according to its action category, thereby preparing the training data for subsequent use in action recognition model training.

As shown in Table I, in TTA20 dataset, serve segments include the first and second shots, while receive segments cover the second and third shots. Discussions with professional table tennis coaches revealed that accurately identifying serve and receive action categories requires considering the opponent's return actions as well, which serve as important comprehensive judgment indicators. Consequently, the serve and receive video segments in the TTA20 dataset also include the subsequent stroke to provide more information for action recognition.

TABLE I. TTA20 Dataset Video Clip Specifications

	Start	End
Serving	Player Serving (1st shot)	4 Frames after Opponent Receiving (2nd shot)
Receiving	4 Frames before Opponent Receiving (2nd shot)	4 Frames after Player Receiving (3rd shot)

B. Table Tennis Playing Status Recognition Model (TTPSRM)

This study establishes the Table Tennis Playing Status Recognition Model (TTPSRM), which is used to identify the action status categories of players in table tennis broadcast video. In the training of TTPSRM, player statuses are divided into three action categories: serve, receive, and others. The "others" category includes actions when players are resting, preparing, or not in a playing state.

As shown in the model architecture in Fig. 1, the TTA20 dataset undergoes the data preprocessing, ultimately only the information of the two players as input for the training data. Each training instance is input into the model frame-by-frame, with each frame containing information on both players. The pose detection for each player includes 25 coordinate points, with each skeletal point having x and y coordinates indicating the point's position in the frame. Thus, each frame consists of 100 features, which are input into the first GRU layer of the model. This is followed by two hidden layers, and the output layer provides the predicted probabilities for each action category. The final result outputs only the category with the highest predicted probability.

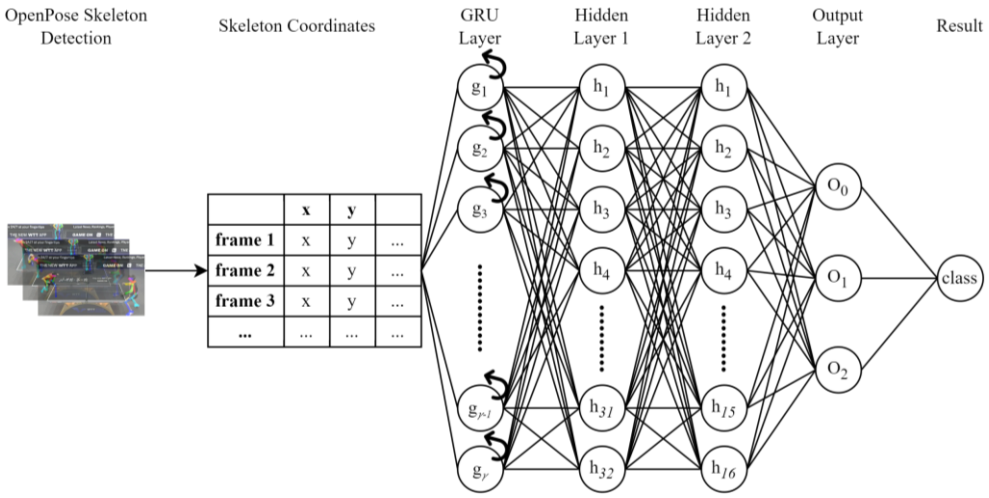

Fig. 1. Table Tennis Playing Status Recognition Model (TTPSRM).

GRU excel at processing sequential data, and each GRU neuron processes the movement changes of the same skeletal points frame-by-frame to capture the relationship between the continuity of skeletal point movements and action categories. Consequently, GRUs effectively capture the temporal features of actions, making them suitable for analyzing skeletal point coordinates extracted by OpenPose for action recognition.

This research delves into the specific application of GRU in action recognition, including the design of the model architecture and the feature extraction method. We designed a GRU network, with each layer containing a different number of neurons to gradually extract higher-level temporal features. Additionally, we introduced an innovative data preprocessing method to ensure the temporal consistency of skeletal point data. Unlike traditional 3D CNN models, which use pixel points and require extensive parameters and computational resources, our GRU-based approach processes skeletal coordinates. This method is more efficient for our task, effectively handling sequential action data. Experimental results show that GRUs offer superior temporal feature extraction and lower computational costs for high-dimensional skeletal data.

C. Full Video Action Timing Detection

The Table Tennis Player Serve and Receive Action Recognition Model (TTPSRM) has been developed for the application of continuous action detection. This model identifies the timing of serve and receive actions throughout the entirety of a table tennis match.

For continuous action recognition of serving and receiving in full match videos, a fixed-size window containing a set number of video frames, with several consecutive frames indicating the occurrence times of the target detected actions within the entire segment.

The continuous detection method begins by capturing the first window from the start of the video. This window comprises several consecutive video frames, which are collectively input into the pre-trained Table Tennis Player Serve and Receive Action Recognition Model (TTPSRM). The model analyzes the motion features within this segment and outputs the classification results for the actions observed.

After classification within the initial window, subsequent windows move forward using a predefined sliding step size. Each new window overlaps the previous one by moving one frame forward, capturing a new consecutive segment of video frames. The frames within the new window are again input as a whole into the TTPSRM for action recognition and result output. Most importantly, when the model outputs a serve or receive classification, it simultaneously records the timing of the event in the video. This timing information is then used for subsequent automated video clipping. This process repeats iteratively until the entire video is analyzed. Therefore, each instance of the sliding window contains a segment of video independently analyzed by the model, generating corresponding action classification results.

This study proposes a continuous action video recognition method based on sliding window technology. By segmenting videos into fixed-size windows and inputting them frame by frame into the action classification model, continuous action recognition in videos is achieved, providing subsequent automated clipping of serve and receive action occurrences in full match videos.

D. Action Clip Automatic Cutting System

Fig. 2 describes the automated video clipping process based on serving timing and receiving timing. The primary input data includes all serving timings and corresponding receiving timings of the video segments, which serve as the basis for video segment clipping. Initially, the system reads a serving timing, processes each serving event sequentially, and identifies the nearest qualifying receiving timing. It checks whether the interval between these timings is less than 4 s to ensure that the selected receiving timing corresponds to the same serve-receive event in the actual video. Upon finding a qualifying receiving timing, the system proceeds to clip the video segment based on the serving and receiving timings, thereby generating a properly clipped serving action segment.

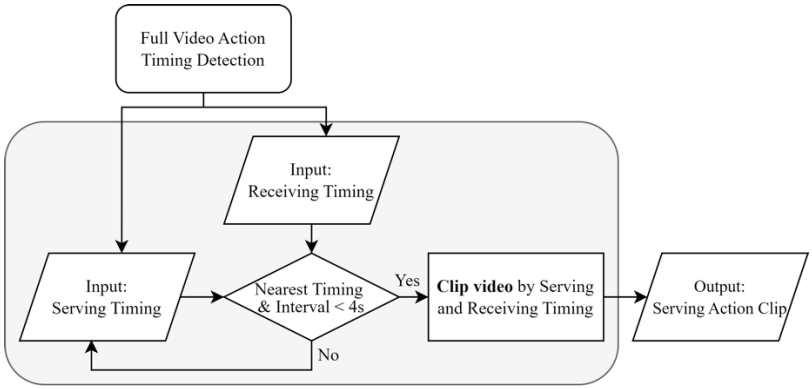

Fig. 2. Action clip automatic cutting system.

Through this process, automated extraction of complete serve action segments from original videos is achieved, facilitating rapid data generation for subsequent analysis and research. Automated processing significantly reduces manual editing labor and enhances data processing efficiency.

IV. Result and Discussion

A. Table Tennis Playing Status Recognition Model (TTPSRM)

The training results of the Table Tennis Player Serve Recognition Model (TTPSRM-Serving) are shown in Tables II and III. TTPSRM-Serve is specifically developed for detecting serve actions, and thus the model training is configured with a window size of 35 and 100 epochs, based on the duration of serve actions. The parameters yielding the highest validation set accuracy during training are selected for final testing data evaluation.

Across the evaluation metrics of F1-Score, Recall, and Precision, the model achieves approximately 0.985. In the confusion matrix from the testing data, most classifications for the three categories are accurate, with only 32 instances of non-serve actions being misclassified as serves. The overall model accuracy stands at approximately 98.4%.

TABLE II. TTPSRM-Serve Performance Evaluation

Metrics	Performance
F1-Score	0.984
Recall	0.984
Precision	0.985

TABLE III. TTPSRM-Serve Confusion Matrix

		Predict		
		Serving	Receiving	Other
Real	Serving	1595	6	1
	Receiving	20	652	6
	Other	12	0	601

The training results of the Table Tennis Player Serve Recognition Model (TTPSRM-Receive) are shown in Tables IV and V. TTPSRM-Receive is specifically developed for detecting receive actions, and thus the model training is configured with a window size of 15 and 100 epochs, based on the duration of receive actions. The parameters yielding the highest validation set accuracy during training are selected for final testing data evaluation.

Across the evaluation metrics of F1-Score, Recall, and Precision, the model achieves approximately 0.967. In the confusion matrix from the testing data, most classifications for the three categories are accurate. However, there are 37 instances where receive actions are incorrectly classified as serves, and 45 instances where serve actions are incorrectly classified as receives. The overall model accuracy stands at approximately 96.6%.

TABLE IV. TTPSRM-Receive Performance Evaluation

Metrics	Performance
F1-Score	0.967
Recall	0.966
Precision	0.967

TABLE V. TTPSRM-Receive Performance Evaluation

		Predict		
		Serving	Receiving	Other
Real	Serving	1550	45	0
	Receiving	37	632	2
	Other	9	4	616

B. Accuracy Assessment—Video Automatically Clipping

This section describes the practical application of the developed automatic video clipping system in live table tennis broadcasts. The evaluation assesses the discrepancies between manually marked points and those determined automatically by the computer, and also examines whether the system missed detecting any playing segments. Following the evaluation methodology outlined in the research, differences in seconds between automatic and manual clipping are calculated, along with the accuracy of serve segment recognition.

C. Serving Action Recognition Accuracy

This study used video from the 2022 International Table Tennis Federation (ITTF) competition held in Mexico to test the automatic clipping system. The match featured players Fan Zhendong and Xu Xin. In the first game of this match, there were a total of 20 serves performed by both players. Following automatic recognition by the clipping model, one starting point was undetected, and there were

five instances where the ending points were not detected. If either the start or end point was not accurately detected, it resulted in the segment not being successfully clipped. Consequently, according to Eq. (1), out of the 20 serving actions, the computer automatically identified and clipped 14 instances, achieving a serve recognition accuracy of 70%.

$$\left(\frac{14}{20}\right) \times 100\% = 70\% \quad (1)$$

D. Timestamp Clipping Accuracy

By comparing the manually marked timings with the computer-generated clipping results, we can evaluate the accuracy of the automatic video clipping. Calculating the difference in seconds between these points provides a measure of this accuracy. The result is shown in Table VI.

For the start points, the computer detected a total of 19 instances. Taking the absolute values of the errors for these 19 instances and averaging them, the final average error in start points is 0.547 s.

Regarding the detection of end points, the computer identified 15 instances. Similarly, taking the absolute values of the errors for these 15 instances and averaging them, the final average error in end points is 0.287 s.

TABLE VI. TIMESTAMP CLIPPING ACCURACY

	Error
Start	0.547 (s)
End	0.287 (s)

V. CONCLUSION

The primary aim of this research is to develop an automated system for extracting and recognizing serve and receive actions from full table tennis match videos. This innovative system significantly reduces the need for human resources by automating the video clipping process, rapidly generating training data, and addressing the large dataset requirements of machine learning. By enhancing the speed and efficiency of data collection, this system improves the accuracy of action recognition models for table tennis, representing a substantial advancement in sports video analysis.

The proposed system, Table Tennis Playing Status Recognition Model (TTPSRM), has shown significant efficacy in accurately identifying players' action statuses, enabling subsequent automatic video clipping. The overall serve recognition accuracy of approximately 98.4% and a receive recognition accuracy of about 96.6%. These results demonstrate that TTPSRM can accurately identify players' statuses in match video, providing valuable data for subsequent systems.

For the performance evaluation of the automated video segmentation system, this research uses videos from the 2022 ITTF competition for testing. The system can automatically clip 70% of the serve action clips, with an average start point error of 0.547 s and an end point error of 0.287 s. Although there are slight errors, the accuracy remains very high and closely matches the results of manual clipping. Some serve action segments were not clipped, but the system can already automatically capture most serve segments, making it suitable for practical data processing workflows.

This research tackles data scarcity by automating the generation of large training datasets, enhancing action recognition accuracy in table tennis. The system's automated clipping of action segments lays the groundwork for applying semi-supervised learning to improve recognition models.

To address the issue of missed serve action segments, future work will focus on improving the system's detection capabilities. Some serve actions are missed due to camera transitions or less distinct serve motions by players. The research aims to develop enhanced algorithms to better handle these special cases. By refining the detection mechanisms, the overall accuracy of the automated video clipping system is expected to increase, ensuring comprehensive data collection for training deep learning models.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Zhong-Kai Wei conducted data preprocessing, model training, organized the result data, and drafted the manuscript. Jieh-Ren Chang provided research methodology guidance, supervised the study, and contributed to the integration of the research content. All authors have approved the final version of the manuscript.

REFERENCES

[1] A. Rossi, L. Pappalardo, and P. Cintia, "A narrative review for a machine learning application in sports: An example based on injury forecasting in soccer," *Sports*, vol. 10, no. 1, 5, Jan. 2022.

[2] R. T. Jewell, M. Konjer, H. E. Meier, and K. Wedeking, "Artificial intelligence and machine learning in sport research: An introduction for non-data scientists," *Frontiers in Sports*, 2022.

[3] A. Sen, S. M. M. Hossain, R. Uddin, K. Deb, and K. H. Jo, "Sequence recognition of indoor tennis actions using transfer learning and long short-term memory," in *Proc. the 28th International Workshop on Frontiers of Computer Vision, IW-FCV 2022*, Hiroshima, Japan, 2022.

[4] C. Tian, V. de Silva, M. Caine, and S. Swanson, "Use of machine learning to automate the identification of basketball strategies using whole team player tracking data," *Appl. Sci.*, vol. 10, no. 1, 24, 2020.

[5] Y.-C. Li, C.-T. Chang, C.-C. Cheng, and Y.-L. Huang, "Baseball swing pose estimation using OpenPose," in *Proc. the 2021 IEEE International Conference on Robotics, Automation and Artificial Intelligence (RAAI)*, 2021.

[6] C.-J. Wang, "3D-convolutional neural network for table tennis serving and receiving action recognition," Master thesis, Department of Electronic Engineering, National Ilan University, July 2023.

[7] A. Hestness, S. Narang, N. Ardalani, G. Diamos, H. Jun, H. Kianinejad, M. Patwary, Y. Yang, and Y. Zhou, "Revisiting unreasonable effectiveness of data in deep learning era," arXiv preprint, arXiv:1707.02968, 2017.

[8] T. Wu. (2022). Camera-Based Table Tennis Posture Analysis. GitHub. [Online]. Available: https://github.com/wutonytt/Camera-Based-Table-Tennis-Posture-Analysis

[9] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," arXiv preprint, arXiv:1412.3555, Dec. 2014.

[10] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "A comparison of LSTM and GRU networks for learning symbolic sequences" arXiv preprint, arXiv:2107.02248, 2021.
[11] M. E. Karim, M. Foysal, and S. Das, *Stock Price Prediction Using Bi-LSTM and GRU-Based Hybrid Deep Learning Approach*, *Springer*, 2023.
[12] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," arXiv preprint, arXiv:1406.1078, 2014.
[13] A. Bharathi, R. Sanku, M. Sridevi *et al*., "Real-time human action prediction using pose estimation with attention-based LSTM network," *Signal, Image and Video Processing*, 2024. https://doi.org/10.1007/s11760-023-02987-0
[14] C.-B. Lin, Z. Dong, W.-K. Kuan, and Y.-F. Huang, "A framework for fall detection based on openpose skeleton and LSTM/GRU models," *Applied Sciences*, vol. 11, no. 2, 329, 2021.
[15] Z. Yan, J. Li, P. Liu, and X. Zhu, "DG-STGCN: Dynamic spatial-temporal modeling for skeleton-based action recognition," arXiv preprint, arXiv:2210.05895, 2022.
[16] M. E. Karim, M. Foysal, and S. Das, "Bi-GRU-attention enhanced unsupervised network for skeleton-based action recognition," *Sensors*, vol. 22, no. 3, 762, 2022.