Recognizing a Sequence of Events from Tennis Video Clips: Addressing Timestep Identification and Subtle Class Differences

Zhaoyu Liu
National University of Singapore
Singapore, Singapore
e0253678@u.nus.edu

Jingyu Guo
National University of Singapore
Singapore, Singapore
guojingyu@u.nus.edu

Mo Wang
National University of Singapore
Singapore, Singapore
e0439309@u.nus.edu

Ruicong Wang
National University of Singapore
Singapore, Singapore
wangruicong@u.nus.edu

Kan Jiang
National University of Singapore
Singapore, Singapore
jiangkan.sg@gmail.com

Jin Song Dong
National University of Singapore
Singapore, Singapore
dongjs@comp.nus.edu.sg

***Abstract*—Detecting temporally precise and fine-grained events from tennis videos is important in automatic video annotation. This paper addresses the challenges of recognizing a sequence of events from tennis video clips, focusing on accurate timestep identification and distinguishing subtle class differences. We propose a novel but simple end-to-end event detection network to accurately detect and identify the key events, which can be trained on a single GPU. We demonstrate that our model outperforms the existing baselines on our fine-grained tennis event dataset. The research contributes to the development of tennis video analytics and has broader implications in other sports domains.**

***Index Terms*—precise event detection, tennis, video analytics, sports analytics**

I. Introduction

The analysis of sports videos plays a crucial role in enhancing athletic performance, refining coaching strategies, and providing valuable insights into player dynamics. In the realm of tennis, the ability to recognize and understand a sequence of events (e.g., players hitting the ball and the ball bouncing on the court) from video clips has gained significant interest due to its potential to empower players, coaches, and analysts alike. However, achieving accurate event detection in tennis videos presents unique challenges that demand innovative solutions.

This paper focuses on two prominent challenges encountered in recognizing a sequence of events from tennis video clips. Firstly, accurate identification of the timesteps at which different events occur is essential. Tennis is a fast-paced sport characterized by swift player movements, ball trajectories, and racket actions, making it vital to precisely pinpoint the temporal occurrence of various activities. Accurate timestep identification forms the foundation for subsequent analysis and enables the extraction of meaningful insights for performance evaluation and strategic decision-making.

Secondly, distinguishing between different classes of events in tennis can be exceptionally challenging due to the subtle differences that exist. For instance, discriminating between a forehand and backhand shot can be perplexing since the overall player body movement appears similar. The distinguishing factor often lies in a specific part of the player's body, occupying only a small portion of the image. Detecting and discerning such nuanced differences is critical for capturing the intricacies of tennis techniques and enabling comprehensive action recognition systems.

Addressing these challenges necessitates the development and deployment of advanced computer vision techniques and machine learning algorithms. Recent advancements in deep learning, specifically Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have shown great promise in addressing complex recognition tasks. These techniques have the potential to capture both temporal and spatial dependencies, enabling the accurate identification of timesteps and subtle class differences in tennis actions.

In this paper, we propose a novel end-to-end approach to recognize a sequence of events from tennis video clips, incorporating state-of-the-art video understanding techniques and deep learning algorithms. Our method seeks to address the challenges of accurately identifying timesteps and discerning subtle differences between different action classes. We present an in-depth analysis of our proposed methodology, experimentally evaluate its performance using a comprehensive dataset of tennis videos, and compare it with existing approaches.

The contributions of this research will not only aid in the development of advanced event detection for tennis but also pave the way for broader applications in other sports domains. By overcoming the challenges of timestep identification and subtle class differences, our work aims to provide valuable insights into tennis techniques, foster player development, and enhance the overall understanding of the game.

The remainder of this paper is organized as follows: Section 2 provides a comprehensive review of related work in the field

of action recognition and highlights the gaps our research aims to address. Section 3 details our proposed methodology, outlining the key components and techniques employed. Section 4 presents the experimental setup and evaluation results. Finally, Section 5 concludes the paper and discusses potential future directions for further research in this domain.

II. Related Work

In this section, we present existing works related to precise action detection from videos.

A. Action recognition

Action recognition is an important task in video understanding that takes one video clip as input to classify the performed action. Models, such as traditional Convolutional Neural Networks (CNNs), SlowFast architecture, and Transformers perform differently for various tasks. It is challenging to determine the definitive best model for tennis action recognition. This paper focuses on investigating the performance of these models.

Convolutional 3D (C3D) Network [1], is a network that extends 2D image models to the spatiotemporal domain, treating spatial and temporal dimensions similarly. Considering the input of multiple frames, 3D convolution would likely work better compared to 2D convolution. This is because, in 2D convolution, each frame in the input is treated independently, while in 3D convolution, the temporal relationship between the frames is reflected in the output, and the information related to the time dimension is kept. The proposed architecture takes in video clips, which go through several convolution and pooling layers to produce the intermediate output. The intermediate output is passed through 2 fully connected layers and a softmax output layer to obtain the C3D features. For the action recognition task, the C3D features are passed into a linear Support Vector Machine for classification, and the performance is better compared to existing models.

SlowFast [2] is a two-stream architecture that handles the temporal dimension uniquely and consists of two pathways with different frame rates. It is designed to draw an analogy with the biological Parvo- and Magnocellular systems. The Slow pathway has a larger temporal stride and more channels. It can learn spatial semantics effectively because information like the player, tennis racket, and clothing colors would not change rapidly within a short duration. The Fast pathway, by contrast, operates with a smaller temporal stride and fewer channels, enabling it to capture swing details within several frames while disregarding semantic information, making it lightweight. Lateral connections fuse these two pathways and facilitate information transfer. SlowFast outperforms spatiotemporal filtering and optical flow models while possessing a lightweight architecture.

Researchers have recently shifted their focus from CNNs to Transformers in video understanding because of the attention mechanism's ability to process global information over large distances. Building through the spatiotemporal adaptation of the Swin Transformer [3], which was designed for image understanding tasks, incorporated inductive bias for spatial locality, as well as for hierarchy and translation invariance, Video Swin Transformer [4] extends its capability to process video data. It extends the scope of local attention computation to the spatiotemporal domain instead of the original spatial domain. The original paper produces 4 different versions for different channel numbers of hidden layers in the first stage and different hyperparameters, Swin-T, Swin-S, Swin-B, Swin-L. The experiment results show that Swin-L with a larger spatial resolution of 384*384 outperforms the state-of-the-art model on the Kinetics-400 and Kinetics-600 datasets.

B. Temporal action detection

Temporal action detection (TAD) is a sophisticated technique aiming to achieve dual objectives: predicting the semantic label and the precise temporal interval of every action instance within an untrimmed video [5]. In the domain of sports video analytics, TAD plays a pivotal role by enabling accurate detection and categorization of crucial action moments.

Liu et al. [6] focused their efforts on detecting hitting moments in badminton. They employed a GRU-based recurrent network that integrated essential features like court position, 2D shuttlecock position, and both players' positions and poses. However, their approach relied on a two-step process, necessitating additional detection models to obtain features. This introduced noise during the learning process and impacted the system's performance.

In contrast, Hong et al. [7] present an innovative end-to-end event detection model named *E2E-Spot*, which utilizes the Temporal Shift Model (TSM) [8] as the local spatial-temporal feature extractor and a bi-directional GRU for long-term temporal reasoning. This model leverages video clips as inputs and demonstrates precise identification of key events in various sports, including tennis, encompassing critical actions like ball-hitting and bouncing events. However, their study failed to consider different shot types (i.e., forehand or backhand), and the model's architecture proved unsuitable for handling more fine-grained event classes. Furthermore, the use of low-resolution input images posed challenges for precise event detection.

In our research, we seek to address these limitations by introducing an advanced and comprehensive approach to TAD. Our method aims to precisely detect 9 different events while taking into account diverse shot types. By leveraging a more refined model architecture and higher-resolution input images, we strive to achieve greater accuracy and reliability in sports action detection.

III. The Proposed Approach

We define the precise action detection as follows: for a given video clip with N frames, we aim to perform the prediction on each frame, which belongs to one of K event classes or a background class indicating no event was detected. We define 8 different event classes: near/far court player serve ball contact, near/far court player forehand swing ball contact, near/far court player backhand swing ball contact, and near/far

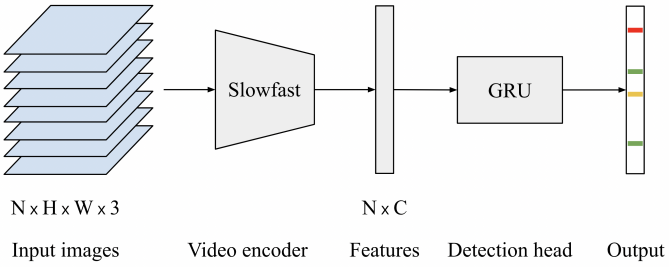

Fig. 1. End-to-end event detection model architecture.

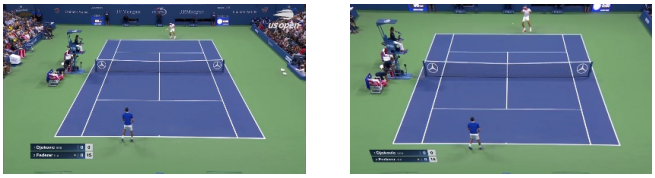

Fig. 2. An example of a frame before and after perspective transformation.

court ball bounce. A prediction is considered correct if its timestamp falls within δ frames of a labeled ground truth and has the same event class.

We propose an end-to-end model for precise event detection. The model architecture is shown in Fig. 1. We take a video clip of size N as input, it will first go through a SlowFast video encoder to obtain the spatial-temporal feature representation for each frame. After that, we use bi-directional GRU as the detection head to perform dense classification on each frame and output the corresponding class scores. In the following sections, we will elaborate on each of the model components.

A. Input data preprocessing

Due to the filming angle, the player and the ball at the far court are usually much smaller and more blurry than at the near court. Therefore, we perform the perspective transformation on each frame as shown in Fig. 2 such that the sizes of both players are similar. Frames are then resized to 224 pixels in height and cropped to 224 × 224 for efficient computation.

B. Video encoder

We select the SlowFast (SF) network as our video encoder. This architecture comprises both a slow pathway and a fast pathway, each processing sparsely and densely sampled video frames, respectively. The fast pathway has a reduced number of channels compared to the slow pathway, allowing it to efficiently capture motion information. This motion information is then progressively fused with the slow pathway at different stages of processing.

We experiment on two different variants for the SF networks with various frame lengths and sample rates: "SlowFast 4×16, R50" and "SlowFast 8×8, R50". Taking "SlowFast 4×16, R50" as an example, given an input clip of N frames, the fast and the slow pathway sample N and $N/8$ frames respectively. We resize the output features of the two pathways to the same length of N and concatenate them into one. The number of channels for each frame is 2,304 (2,048 for the slow pathway and 256 for the fast pathway).

Compared with other video encoders such as TSM [8] and C3D [1], SF captures both sparse and dense temporal information with different step sizes, it is more suitable for the purpose of detecting events that require both precise temporal spotting with subtle visual differences and accurate event classification that requires a larger stride size.

C. Detection head

In order to obtain long-term temporal information, we use a 1-layer bidirectional Gated Recurrent Unit (GRU) network, which processes the dense per-frame features. Finally, we apply a fully connected layer and softmax on the GRU outputs to make a per-frame $K + 1$ way prediction (including 1 background class).

D. Per-frame cross-entropy loss

For a video clip of N frames, our model will output N predictions of size $K + 1$ considering the background class. The prediction can be denoted as $(\hat{y_1}, \hat{y_2}, ..., \hat{y_N})$. The ground-truth label for each frame is denoted as $y_t \in \{c_1, c_2, ..., c_k, c_{background}\}$. The loss function is computed as:

$$loss = \sum_{t=1}^{N} CrossEntropy(\hat{y_t}, y_t)$$

E. Implementation details

Our model is trained on randomly sampled 64-frame-long clips, with standard data augmentations like cropping, jitter, and mixup [9] applied. The frames are resized to a height of 224 pixels and cropped to 224 × 224 pixels. We employ AdamW optimization [10] with LR annealing [11]. To address the imbalance caused by the rarity of precise events ($< 3\%$ of frames), foreground class loss weights are increased by a factor of 5 compared to the background.

During testing, data augmentation is turned off, and clips overlap by 50%, with per-frame predictions averaged. For the conversion of per-frame class scores to spotting predictions, frames are ranked by their predicted scores for each class.

IV. Evaluation

A. Dataset

Our dataset is an extension of the dataset published by [7] for the tennis events spotting task. 28 broadcast videos from different matches have been labeled with frame-accurate events, including "serve ball contact", "forehand ball contact", "backhand ball contact", and "ball bounce" (each divided by near- and far-court). 19 videos are used for training and validation and 9 are dedicated to testing. Tab. I shows a detailed description of the dataset. To weaken the imbalance between background/ foreground events, we randomly sample background events within a rally timestamp, with a comparable number to foreground events.

TABLE I
EVENT CLASSES AND THEIR COUNTS FOR TENNIS DATASET.

Event class	Train	Val	Test
Near-court serve	672	238	780
Near-court forehand swing	1,209	380	2,148
Near-court backhand swing	991	329	1,988
Near-court bounce	2,607	871	4,650
Far-court serve	658	200	799
Far-court forehand swing	1,197	364	2,024
Far-court backhand swing	1,022	393	2,123
Far-court bounce	2,620	867	4,662

TABLE II
PERFORMANCES OF DIFFERENT MODELS.

Models	$\delta = 1$	$\delta = 2$
Sliding window method		
VideoSwin_Tiny	68.96%	83.79%
SlowFast-4x16	83.00%	85.91%
End-to-end per-frame classification method		
E2E-Spot (w/o persp)	90.67%	92.46%
E2E-Spot (w persp)	91.98%	93.47%
E2E-SlowFast-4x16 (w persp)	92.42%	94.24%
E2E-SlowFast-8x8 (w persp)	**95.03%**	**96.52%**

B. Evaluation metric

For evaluating our approach, we adopt the mean Average Precision within a tolerance of δ frames (mAP@δ) as our primary metric. To ensure precise event spotting, we compute the Average Precision for each foreground class and subsequently calculate their mean. To emphasize the precision of event detection, we report the mAP scores at 1 and 2 frames, thereby assessing the accuracy of our method in precisely localizing events in the video sequences.

C. Baselines

In our evaluation, we compare the performance of our method against baselines derived from both action recognition and temporal action detection domains.

1) Action Recognition with Sliding Window: For action recognition, we implement a sliding window approach using 15 frames, training the model to predict the event that happens in the middle frame. To provide a comprehensive comparison, we utilize two state-of-the-art action recognition models: SlowFast and Video Swin Transformer. These baselines serve as essential references for evaluating the effectiveness of our proposed method.

2) E2E temporal action detection: For temporal action detection, we adopt E2E-Spot [7] as our baseline. Additionally, we investigate the impact of perspective transformation on input data pre-processing and compare the performance with and without this technique.

D. Performance

As shown in Table II, the end-to-end per-frame classification method shows a better performance than the sliding window method in the context of temporal action detection. The reason could be that the window size for the sliding window method is too large to identify the precise moment of events happening.

For the end-to-end dense classification method, the models with perspective transformation exhibit better performances because both of the players are larger and of similar sizes. Our proposed method with SlowFast network as video encoder performs better than the E2E-Spot at $\delta = 1$ and 2. One of the possible reasons is that the local spatial-temporal visual information captured by E2E-Spot using TSM is too short to identify fine-grained actions (e.g., forehand and backhand) because they usually involve longer frame lengths. While the SlowFast structure enables the combination of various stride sizes, which can not only precisely spot the event moments but also accurately identify the event types. Among all the models, the SlowFast 8 $\times$ 8 has the best performance with 95.03% and 96.52% mAP on $\delta = 1$ and 2, respectively.

V. CONCLUSION AND FUTURE WORK

In this paper, we have presented an end-to-end event detection model to accurately detect critical event moments and types in broadcast tennis videos. Compared with previous work, we aim to detect more fine-grained event classes by including action types. Besides, we have proposed a better model architecture by applying SlowFast as our video encoder which is proved more suitable for our purpose of detecting events that require both precise temporal spotting and accurate event classification that requires a larger stride size. Moreover, we have proposed an input preprocessing step to further boost the performance while not introducing additional model complexity. Our method achieves state-of-the-art performance on the public fine-grained tennis event detection dataset and can be easily adapted to other sports such as badminton, table tennis, and soccer.

Promising research directions involve methodological enhancements, such as developing improved video encoding architectures (e.g., based on visual transformers), training methodologies, head architectures, and losses that benefit from end-to-end learning. Additionally, a focus on more fine-grained event classes can help include spatial information, such as player positions at hitting moments and ball-bouncing locations, while also incorporating a wider range of action types, such as volley, topspin, and slice. To enrich the dataset's utility, we aim to expand it by incorporating more broadcast videos and college/junior matches, benefiting players across all levels.

REFERENCES

[1] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[2] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6202–6211.

[3] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[4] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 3202–3211.
[5] X. Liu, S. Bai, and X. Bai, "An empirical study of end-to-end temporal action detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 010–20 019.
[6] P. Liu and J.-H. Wang, "Monotrack: Shuttle trajectory reconstruction from monocular badminton video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3513–3522.
[7] J. Hong, H. Zhang, M. Gharbi, M. Fisher, and K. Fatahalian, "Spotting temporally precise, fine-grained events in video," in *European Conference on Computer Vision*. Springer, 2022, pp. 33–51.
[8] J. Lin, C. Gan, and S. Han, "Tsm: Temporal shift module for efficient video understanding," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 7083–7093.
[9] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.
[10] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.
[11] ——, "Sgdr: Stochastic gradient descent with warm restarts," in *International Conference on Learning Representations*, 2016.