

Two-Stream Network and Attention Mechanism for Sports Video Classification in Table tennis

Pengcheng Dong¹, Hongxin Xie¹, Fuqiang Zheng² and Jiande Sun^{1,†}

¹*School of Information Science and Engineering, Shandong Normal University, Jinan, China*

²*School of Physical Education, Shandong Normal University, Jinan, China*

Abstract

Precise recognition of fine-grained actions in sports videos requires robust models proficient in capturing intricate spatiotemporal cues. Our study introduces a novel hybrid framework that combines SlowFast[1] for refined temporal modeling with CBAM[2] for channel and spatial attention. Additionally, TAM[3] integrates sophisticated temporal attention mechanisms within our innovative architecture. Our model aims to elevate the comprehension and identification of intricate actions within high-speed sports videos, with a specific emphasis on table tennis. We validate our proposed framework using the rigorous TTStroke-21 dataset[4, 5], showcasing its superior performance in fine-grained action classification and accurate position detection within table tennis videos. Experimental outcomes vividly demonstrate the efficacy of our hybrid approach in discerning nuanced stroke variations and precisely localizing actions[6], signifying its substantial potential in sports analytics and comprehensive player performance assessment.

1. Introduction

Recent years have witnessed a notable surge in sports video analysis, notably in the nuanced deciphering of intricate movements within table tennis videos[7]. Advanced methodologies have become pivotal in extracting comprehensive insights into player performance, enabling coaches and analysts to refine training strategies and optimize athletes' potential.

Our proposed framework stands as a significant stride in this domain, amalgamating a sophisticated fusion of cutting-edge techniques. Employing SlowFast[1] for temporal modeling assures a profound comprehension of temporal dynamics, empowering our model to discern the intrinsic rapid stroke variations prevalent in the realm of table tennis. Moreover, the CBAM[2] dynamically recalibrates channel and spatial information, while the TAM[3] refines temporal representations. Their collective effect significantly boosts our model's precision in recognizing fine-grained actions within video sequences.

This choice of model was propelled by SlowFast's capability to harmoniously combine spatial and temporal cues in sports videos, particularly suited for the fast-paced nature of table tennis. The incorporation of SlowFast serves as a cornerstone in our model, offering nuanced insights into temporal dynamics crucial for recognizing and categorizing fine-grained actions within the context of table tennis. Therefore, its selection was grounded in its capacity to comprehend the rapid and intricate motions intrinsic to this high-speed sport, aiming to substantially enhance sports analytics methodologies and further our understanding of athlete performance in dynamic sporting environments.


MediaEval'23: Multimedia Evaluation Workshop, February 1–2, 2024, Amsterdam, The Netherlands and Online

[†]The corresponding author.

✉ 2022317067@stu.sdn.u.edu.cn (P. Dong); 2502174276@qq.com (H. Xie); zhengfuqiang1981@sina.com (F. Zheng); jiandesun@hotmail.com (J. Sun)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

2. Approach

Table tennis, as a fast-paced and dynamic sport, presents unique challenges for action recognition and precise position detection. To address these challenges, advanced video analysis techniques utilizing 3D Convolutional Neural Networks (CNNs) [8] have emerged as a promising solution. Within this context, the SlowFast[1] and ResNet3D[9] architectures have demonstrated considerable potential in enhancing fine-grained action classification and position detection in Table tennis videos.

2.1. Integration of Attention Mechanisms

To refine Fine-Grained Action Classification and Position Detection within Table tennis video analysis, our network architecture incorporates attention mechanisms such as Channel Attention Block (CBAM)[2] for channel-specific focus and Spatiotemporal Attention Mechanism (TAM)[3] for temporal insights. Seamlessly integrated, these mechanisms significantly enhance discriminative capabilities, effectively capturing intricate spatiotemporal patterns inherent in Table tennis sequences.

The CBAM[2] module enriches network capability by focusing on insightful channel-wise relationships within feature maps. It adaptively recalibrates feature responses to emphasize salient features while mitigating irrelevant information. The integration of CBAM[2] facilitates the extraction of discriminative spatial features, enhancing fine-grained action classification and precise position detection within Table tennis sequences.

Simultaneously, TAM [3] adeptly captures long-range dependencies and temporal relationships in Table tennis videos. Operating through attention mechanisms across temporal dimensions, TAM[3] empowers the model to highlight crucial temporal information, significantly contributing to the precision of action recognition and position detection. It effectively filters out redundant frames and accentuates subtle temporal dynamics during gameplay.

2.2. SlowFast Networks

The SlowFast[1] architecture adeptly processes spatial and temporal information via dual pathways, enabling a thorough analysis of motion dynamics in Table tennis videos. While the slow pathway meticulously captures intricate spatial details, the fast pathway focuses on rapid temporal changes, facilitating the fusion of detailed spatial and dynamic temporal features. This fusion significantly augments the precision of identifying fine-grained actions and accurately detecting player positions during gameplay. Furthermore, the integration of Adaptive Time Attention accentuates critical temporal segments, enhancing the network's proficiency in discerning significant temporal dynamics, thereby refining action recognition and position detection[8]. In summary, the incorporation of SlowFast[1] markedly amplifies the accuracy of Fine-Grained Action Classification and Position Detection in Table tennis videos, showcasing its pivotal role in advancing the landscape of sports video analysis. The model architecture, depicted in Figure 1, showcases the integration of CBAM[2] within the slow branch and TAM within the fast branch. CBAM's[2] placement within the slow branch capitalizes on its capability to extract spatial and channel-related details, given the branch's lower count of image frames but richer channel information. Conversely, in the fast branch characterized by fewer channel details but more image frames, TAM excels in capturing temporal relationships among these frames.

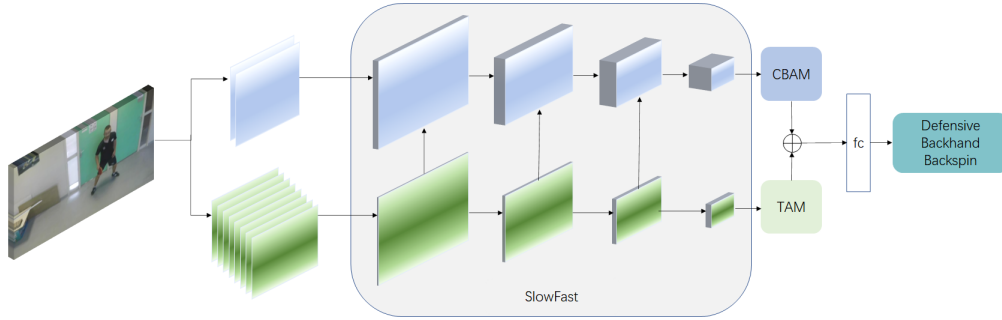


Figure 1: The integration of SlowFast[1] architecture, CBAM[2], and TAM[3] within the framework ensures robust spatiotemporal analysis, enhancing fine-grained action classification and position detection in Table tennis videos.

3. Results and Analysis

This study presents a comprehensive experimental framework for video classification employing the sophisticated architecture termed SlowFast[1]. Our model was benchmarked against Timesformer[10] and SlowFast[1] in experimental comparisons. Timesformer[10] is a neural network model designed for time series data, utilizing a Transformer[11] architecture optimized with time-based attention mechanisms to enhance temporal feature processing. We use the pre-training model from open-mmlab in [12]. The experimental setup involves a batch size of 8 samples, an initial learning rate of $1e-2$, weight decay of $1e-5$, momentum of 0.9, encompassing a training regime spanning 500 epochs. Furthermore, employing the cross-entropy loss function for classification tasks[13], a dynamic learning rate scheduler, based on validation set performance, ensures continual performance improvement.

Table 1

Slowfast[1], timesformer[10] and our model experiment results, these results are the running results on our local side.

model	top1(%)	top5(%)
timesformer	82.17	98.70
slowfast	81.30	99.13
slowfast+CBAM	84.28	98.65
slowfast+CBAM+TAM	87.39	98.69

In Table 1, the accuracy of Timesformer is reported as 82.17%. Subsequent ablation experiments were carried out to substantiate the superiority of our model. Accuracy was recorded at 82.17% when employing only SlowFast, which increased to 84.28% upon integrating CBAM with SlowFast. Further enhancement to 87.39% was achieved by combining both CBAM and TAM with SlowFast. Additionally, comparative analysis against the Baseline model demonstrated an accuracy of 74.6% for our model, as illustrated in Table 2. Our analysis identifies three primary reasons for these errors. Firstly, despite the SlowFast network's capability in capturing spatiotemporal information, it might encounter challenges in discerning very subtle or nuanced actions, especially in highly dynamic sports like table tennis. This limitation could affect the precision of action classification and position detection by not adequately capturing fine-grained details. Secondly, the dual-pathway design of SlowFast introduces a trade-off between capturing spatial details and temporal dynamics. Achieving a balance between these pathways to effectively capture both spatial and temporal information remains a challenge, impacting the model's consistency in discerning fine-grained actions and accurately detecting player positions. Lastly,

within the defensive and offensive datasets, there exists similarity in actions despite the different labels assigned. While defensive data includes "backspin," "block," and "push," offensive data comprises "flip," "hit," and "loop." Despite the different labels, the high similarity in the actions poses challenges for accurate classification.

Table 2

Comparison between our model and Baseline[14] on the test set.

model	Acc(%)
Baseline	86.4
slowfast+CBAM+TAM	74.6

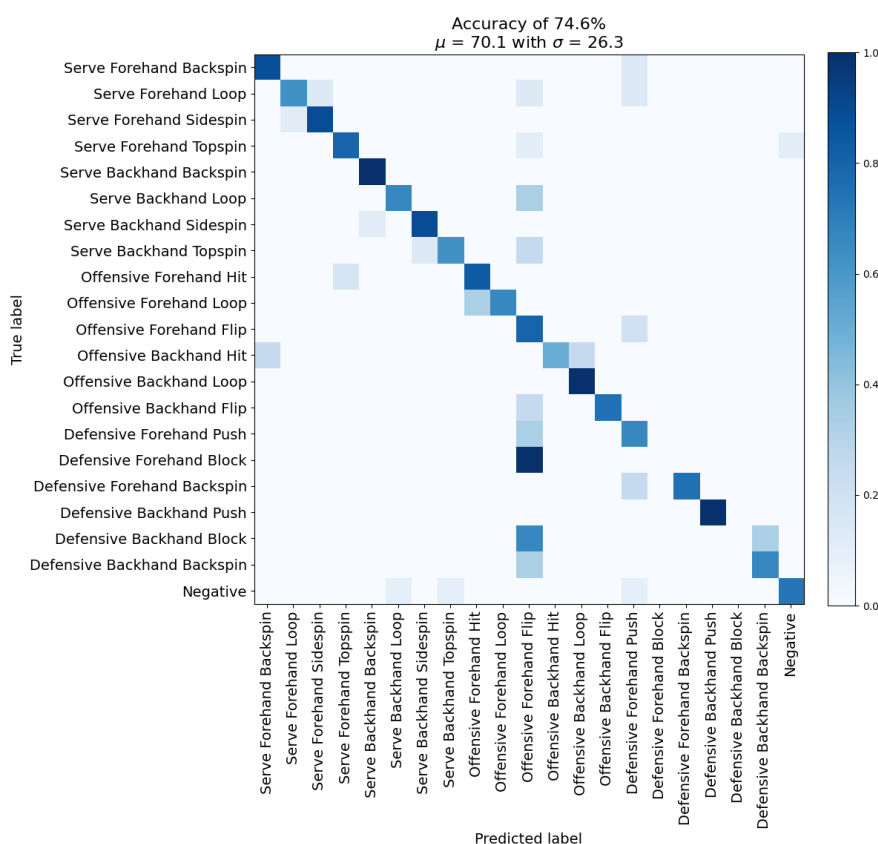


Figure 2: Confusion matrix using all classification combinations on the test set.

4. Discussion and Outlook

Future research endeavors may concentrate on refining the model’s adaptability to diverse scenarios within table tennis matches. Exploring supplementary attention mechanisms or incorporating diverse deep learning architectures holds promise in augmenting comprehension and precision for identifying intricate table tennis movements. Moreover, investigating transfer learning and generalization across diverse sports domains could significantly broaden the scope of applying this technology in sports video analysis.

References

- [1] C. Feichtenhofer, H. Fan, J. Malik, K. He, Slowfast networks for video recognition, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 6202–6211.
- [2] S. Woo, J. Park, J.-Y. Lee, I. S. Kweon, Cbam: Convolutional block attention module, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 3–19.
- [3] P. J. Hu, P. Y. Chau, O. R. L. Sheng, K. Y. Tam, Examining the technology acceptance model using physician acceptance of telemedicine technology, *Journal of management information systems* 16 (1999) 91–112.
- [4] P.-E. Martin, J. Benois-Pineau, R. Péteri, J. Morlier, Fine grained sport action recognition with twin spatio-temporal convolutional neural networks: Application to table tennis, *Multimedia Tools and Applications* 79 (2020) 20429–20447.
- [5] P.-E. Martin, J. Benois-Pineau, R. Péteri, J. Morlier, Sport action recognition with siamese spatio-temporal cnns: Application to table tennis, in: 2018 International Conference on Content-Based Multimedia Indexing (CBMI), IEEE, 2018, pp. 1–6.
- [6] A. Erades, P. Martin, R. V. B. Mansencal, R. Péteri, J. Morlier, S. Duffner, J. Benois-Pineau, Sportsvideo: A multimedia dataset for event and position detection in table tennis and swimming, in: Working Notes Proceedings of the MediaEval 2023 Workshop, Amsterdam, The Netherlands and Online and Online, 1-2 February 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2023.
- [7] H. Li, S. G. Ali, J. Zhang, B. Sheng, P. Li, Y. Jung, J. Wang, P. Yang, P. Lu, K. Muhammad, et al., Video-based table tennis tracking and trajectory prediction using convolutional neural networks, *Fractals* 30 (2022) 2240156.
- [8] S. Ji, W. Xu, M. Yang, K. Yu, 3d convolutional neural networks for human action recognition, *IEEE transactions on pattern analysis and machine intelligence* 35 (2012) 221–231.
- [9] Z. Qiu, T. Yao, T. Mei, Learning spatio-temporal representation with pseudo-3d residual networks, in: proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5533–5541.
- [10] G. Bertasius, H. Wang, L. Torresani, Is space-time attention all you need for video understanding?, in: ICML, volume 2, 2021, p. 4.
- [11] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, Y. Wang, Transformer in transformer, *Advances in Neural Information Processing Systems* 34 (2021) 15908–15919.
- [12] M. Contributors, Openmmlab’s next generation video understanding toolbox and benchmark, <https://github.com/open-mmlab/mmaaction2>, 2020.
- [13] L. Hui, M. Belkin, Evaluation of neural architectures trained with square loss vs cross-entropy in classification tasks, *arXiv preprint arXiv:2006.07322* (2020).
- [14] P. Martin, Baseline method for the sport task of mediaeval 2023 3d cnns using attention mechanisms for table tennis stoke detection and classification., in: Working Notes Proceedings of the MediaEval 2023 Workshop, Amsterdam, The Netherlands and Online and Online, 1-2 February 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2023.