Multi-Task Temporal and Spatial Networks for High-Precision Event Spotting in Volleyball Videos

Hoang Quoc Nguyen[1][2] [a], Second Author Name[1] [b] and Third Author Name[2] [c]

[1]*Korea Institute of Science and Technology, Seoul, Republic of Korea*

[2]*University of Science and Technology, Daejeon, Republic of Korea*

523503, second_author}@kist.re.kr, third_author@dc.mu.edu

Abstract: Understanding the precise timing and location of events is crucial for analyzing sports videos, especially in fast-paced sports like volleyball. We introduce a new task: high-precision spatial-temporal event spotting, which aims to detect both when and where key actions occur. To support this, we present the KOVO Volleyball Event Dataset, featuring 947 rally videos, and 5,935 events, annotated for both temporal and spatial localization. Our best model achieves a combined mAP of 85.46 across various temporal and spatial thresholds. Notably, we find that incorporating spatial predictions enhances temporal mAP by 5.89 points, underscoring the synergy between spatial and temporal analysis. To the best of our knowledge, this is the first work addressing this task, establishing a strong baseline for future research in spatial-temporal event spotting.

1 INTRODUCTION

Video understanding has emerged as a cornerstone in computer vision, offering valuable insights into dynamic scenes for applications such as sports analytics, surveillance, and autonomous systems. Within this field, various tasks have been defined to interpret actions over time. *Temporal Action Detection (TAD)* focuses on pinpointing time intervals where specific actions occur within untrimmed videos, while *Temporal Action Segmentation (TAS)* aims to divide videos into continuous sequences of actions. Complementing these is the task of *Action Spotting*, which identifies the precise frames that capture key events, requiring models to discern subtle temporal differences and visually similar frames (Hong et al., 2022).

Recent advancements in action spotting, such as *T-DEED* (Xarles et al., 2024) and *E2E-Spot* (Hong et al., 2022), have demonstrated the ability of models to achieve frame-level precision in fast-paced events using deep learning architectures. Datasets like *FigureSkating* (Hong et al., 2021) and *FineDiving* (Xu et al., 2022) have been pivotal in advancing action spotting, emphasizing the importance of precise

---
[a] https://orcid.org/0009-0002-2004-9285

[b] https://orcid.org/0000-0000-0000-0000

[c] https://orcid.org/0000-0000-0000-0000

temporal detection in sports with individual athletes. However, these datasets are tailored to specific sports and do not capture the complexity and rapid dynamics of team-based, high-speed sports, such as volleyball.

In volleyball, rapid play transitions occur within specific areas of the court, making precise spatial localization as important as temporal accuracy. To address this, we introduce the new task of *high-precision spatial-temporal event spotting*, designed to detect both the exact timing and spatial location of key events. Unlike conventional action spotting, this task provides richer insights into player positioning and movement patterns, crucial for analyzing volleyball gameplay.

In other sports, datasets like *SoccerNet-v2* (Deliege et al., 2021) have pushed the boundaries of action spotting through rich temporal and spatial annotations, significantly advancing model capabilities. Yet, no equivalent dataset exists for volleyball, a sport characterized by its rapid exchanges and the need for precise localization of actions. To fill this gap, we introduce the *KOVO Event Dataset*, comprising 947 rally videos, 890,797 frames, and 5,935 annotated key actions. This dataset offers granular annotations for both temporal and spatial event localization, making it a valuable resource for developing models that capture the intricacies of volleyball.

Our contributions are threefold. First, we intro-

duce the new task of high-precision spatial-temporal event spotting, specifically tailored for the dynamics of volleyball. Second, we present the *KOVO Event Dataset*, the first of its kind to include detailed temporal and spatial annotations for volleyball rallies. Third, we propose a multi-task deep learning model that jointly predicts event timing and spatial positions, leveraging this dual focus to achieve improved performance. Notably, incorporating spatial predictions into our model enhances temporal mAP by 5.89 points. Our best model achieves a temporal mAP of 90.59, a spatial mAP of 77.94, and a combined mAP of 85.46, providing a strong baseline for this new task. To the best of our knowledge, this work is the first to explore high-precision spatial-temporal event spotting in volleyball, setting the stage for future research in this area.

2 RELATED WORK

2.1 Video Classification

In video understanding, video classification focuses on predicting a single label for the entire video, unlike event spotting, which demands precise frame-level labeling. This difference leads to distinct challenges: video classification often benefits from sparse frame sampling (Wang et al., 2016), while event spotting requires dense sampling to capture rapid changes in events. Additionally, classification models frequently use techniques like global space-time pooling (Tran et al., 2018) or temporal consensus (Zhou et al., 2018) to derive a video-level prediction, which contrasts with the need for maintaining high temporal resolution in event spotting.

Drawing from insights provided by E2E-Spot, which demonstrated the advantages of end-to-end training without temporal pooling for frame-level precision, our approach adopts RegNet-Y (Radosavovic et al., 2020) combined with GSM (Sudhakaran et al., 2020). RegNet-Y, known for its efficient architecture, paired with GSM for adaptive temporal shifts, offers a robust solution for extracting spatial-temporal features. This combination proved particularly effective for our high-precision event spotting task in volleyball, enabling both temporal accuracy and spatial precision while keeping the process efficient.

2.2 Group Activity Recognition

2.3 Precise Action Spotting

3 DATASET OVERVIEW

3.1 Data Content and Statistics

3.2 Annotation Process

3.3 Dataset Splits

4 PROPOSED METHOD

4.1 Problem Formulation

4.2 Model Architecture

4.2.1 Feature Extractor

4.2.2 Temporal Event Detection

4.2.3 Spatial Event Detection

4.2.4 Multi-Task Learning

4.2.5 Loss Function

5 EXPERIMENTS

5.1 Implementation Details

5.2 Training Strategy

5.3 Evaluation Metrics

6 CONCLUSIONS

Please note that ONLY the files required to compile your paper should be submitted. Previous versions or examples MUST be removed from the compilation directory before submission.

We hope you find the information in this template useful in the preparation of your submission.

ACKNOWLEDGEMENTS

If any, should be placed before the references section without numbering. To do so please use the following command:

REFERENCES

Deliege, A., Cioppa, A., Giancola, S., Seikavandi, M. J., Dueholm, J. V., Nasrollahi, K., Ghanem, B., Moeslund, T. B., and Van Droogenbroeck, M. (2021). Soccernet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4508–4519.

Hong, J., Fisher, M., Gharbi, M., and Fatahalian, K. (2021). Video pose distillation for few-shot, fine-grained sports action recognition. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9234–9243.

Hong, J., Zhang, H., Gharbi, M., Fisher, M., and Fatahalian, K. (2022). Spotting temporally precise, fine-grained events in video.

Radosavovic, I., Kosaraju, R. P., Girshick, R., He, K., and Dollár, P. (2020). Designing network design spaces.

Sudhakaran, S., Escalera, S., and Lanz, O. (2020). Gate-shift networks for video action recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1099–1108.

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., and Paluri, M. (2018). A closer look at spatiotemporal convolutions for action recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6450–6459.

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., and Van Gool, L. (2016). Temporal segment networks: Towards good practices for deep action recognition. In Leibe, B., Matas, J., Sebe, N., and Welling, M., editors, *Computer Vision – ECCV 2016*, pages 20–36, Cham. Springer International Publishing.

Xarles, A., Escalera, S., Moeslund, T. B., and Clapés, A. (2024). T-deed: Temporal-discriminability enhancer encoder-decoder for precise event spotting in sports videos.

Xu, J., Rao, Y., Yu, X., Chen, G., Zhou, J., and Lu, J. (2022). Finediving: A fine-grained dataset for procedure-aware action quality assessment.

Zhou, B., Andonian, A., Oliva, A., and Torralba, A. (2018). Temporal relational reasoning in videos.

APPENDIX

If any, the appendix should appear directly after the references without numbering, and not on a new page. To do so please use the following command: