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. This field encompasses various tasks designed to interpret and analyze actions over time. Among these, *Video Classification* aims to assign a single label to an entire video, providing a broad understanding of the content but often lacking frame-level precision. In contrast, *Temporal Action Localization (TAL)* focuses on identifying time intervals where specific actions occur within untrimmed videos. Complementing these is *Precise Action Spotting (PES)*, which identifies the precise frames that capture key events, requiring models to discern subtle temporal differences and distinguish 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

---

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

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

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

et al., 2022) have been pivotal in advancing action spotting, emphasizing the importance of precise 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, which are 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 cap-

ture the intricacies of volleyball.

Our contributions are threefold:

- **New Task Introduction:** We introduce the task of high-precision spatial-temporal event spotting, specifically tailored for the dynamics of volleyball.

- **Dataset Development:** We present the *KOVO Event Dataset*, the first of its kind to include detailed temporal and spatial annotations for volleyball rallies, aimed at fostering research in this area.

- **Model Development:** 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

Video classification aims to predict a single label for an entire video, in contrast to event spotting, which requires precise frame-level labeling. This distinction introduces unique challenges: video classification can leverage sparse frame sampling (Wang et al., 2016), whereas event spotting demands dense sampling to capture rapid changes in events. Additionally, classification models often employ global space-time pooling (Tran et al., 2018) or temporal consensus (Zhou et al., 2018) to produce video-level predictions, while event spotting necessitates preserving high temporal resolution.

2.2   Temporal Action Localization

Temporal Action Localization (TAL) aims to identify the time intervals when specific actions occur in untrimmed videos, making it ideal for longer actions that are not instantaneous. Unlike video classification, which assigns a single label to an entire video, TAL requires precise start and end times, making it more complex.

TAL methods are typically categorized into two groups: two-stage (Qing et al., 2021; Escorcia et al., 2016) and one-stage (Shi et al., 2023a; Zhang et al., 2022) approaches. Two-stage models generate action proposals before classifying them, while one-stage models directly predict actions and their intervals in a streamlined process. Recent methods, like Action-Former (Zhang et al., 2022) and TriDet (Shi et al., 2023b), leverage advanced architectures, including transformers and feature pyramids, to improve temporal precision across varying action durations. Anchor-free approaches (Yang et al., 2020) have further enhanced flexibility in predicting actions without relying on predefined time windows.

TAL's development has been driven by extensive datasets and benchmarks such as ActivityNet (Heilbron et al., 2015), EPIC-KITCHENS (Damen et al., 2018), and THUMOS Challenge (Idrees et al., 2017), making it a well-explored field for understanding complex, prolonged actions in videos. However, it remains distinct from action spotting, which focuses on identifying brief, precise moments in fast-paced scenarios.

2.3   Precise Event Spotting (PES)

Precise Event Spotting (PES) aims to detect the exact frames of key events in untrimmed videos, making it ideal for fast, critical moments in sports. Unlike Temporal Action Localization (TAL), which spans broader time intervals, PES requires frame-level accuracy, crucial for detailed sports analysis where slight timing shifts can alter game interpretations.

Datasets like *FigureSkating* (Hong et al., 2021) and *FineDiving* (Xu et al., 2022) have advanced PES with frame-level annotations, but focus on simpler, individual sports. Recent methods, such as *T-DEED* (Xarles et al., 2024) and *E2E-Spot* (Hong et al., 2022), enhance precision by refining temporal representations and avoiding pooling, but lack spatial context needed for team sports.

In volleyball, spatial localization is as critical as timing. Understanding event locations is key to analyzing player movements and strategies. Our work addresses this by introducing a new task and dataset for high-precision spatial-temporal event spotting in volleyball, capturing both event timing and location.

Inspired by *E2E-Spot* (Hong et al., 2022), our approach uses RegNet-Y (Radosavovic et al., 2020) with GSM (Sudhakaran et al., 2020) for adaptive temporal shifts. This combination balances efficiency and precision, making it ideal for capturing complex spatial-temporal dynamics in volleyball.

3 DATASET OVERVIEW

3.1 Data Content and Statistics

3.2 Annotation Process

3.3 Dataset Splits

3.4 Release and Access

Due to the large size of the dataset, we are unable to release the full-resolution (1280x720) videos in this paper. However, we have made a resized version (512x288 resolution) along with the corresponding annotations available on Kaggle, totaling 100GB. The dataset can be accessed at [provide link].

4 PROPOSED METHOD

4.1 Problem Formulation

Temporal Spatial Event Spotting (TSES) aims to identify both the precise time and location of events within untrimmed videos. Given a video with N frames $x_1, \ldots, x_N$ and a set of K event classes $c_1, \ldots, c_K$, the goal is to predict a sparse set of frame indices where events occur, as well as the corresponding event class and spatial coordinates. Each prediction is represented as (t, c_t, s_t), where t is the frame index, c_t is the predicted event class, and s_t is the spatial location of the event within that frame. A temporal prediction is considered correct if it falls within a small temporal tolerance σ_f frames of the labeled event and matches the ground-truth class. Similarly, spatial predictions are deemed correct if the distance between the predicted location s_t and the ground-truth event location is within a specified threshold σ_p. For TSES, these tolerances are kept small, requiring precise frame-level accuracy, and it assumes that videos are captured with sufficiently high frame rates.

To perform well on this task, several key requirements must be met:

- **Local Spatial-Temporal Features**: The model should capture subtle visual changes and movements across neighboring frames, essential for distinguishing between similar-looking moments.

- **Long-Term Temporal Context**: A broader temporal window allows the model to understand events in context, such as how player positions or actions evolve before and after a critical moment.

- **Dense Per-Frame Predictions**: For each frame x_t, the model produces a prediction $(\hat{y}_t, \hat{s}_t)$, where $\hat{y}_t \in \mathbb{R}^K$ is a one-hot vector representing the predicted event class probabilities, and $\hat{s}_t \in \mathbb{R}^2$ represents the x, y coordinates of the event within the frame. This ensures frame-level precision for each event.

These requirements call for a robust and end-to-end trainable architecture capable of leveraging both temporal and spatial information. Our approach utilizes a sequence model that integrates local spatial-temporal features.

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

Damen, D., Doughty, H., Farinella, G. M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J.,

Perrett, T., Price, W., and Wray, M. (2018). Scaling egocentric vision: The epic-kitchens dataset.

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.

Escorcia, V., Caba Heilbron, F., Niebles, J. C., and Ghanem, B. (2016). Daps: Deep action proposals for action understanding. In Leibe, B., Matas, J., Sebe, N., and Welling, M., editors, *Computer Vision – ECCV 2016*, pages 768–784, Cham. Springer International Publishing.

Heilbron, F., Escorcia, V., Ghanem, B., and Niebles, J. C. (2015). Activitynet: A large-scale video benchmark for human activity understanding.

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.

Idrees, H., Zamir, A. R., Jiang, Y.-G., Gorban, A., Laptev, I., Sukthankar, R., and Shah, M. (2017). The thumos challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding*, 155:1–23.

Qing, Z., Su, H., Gan, W., Wang, D., Wu, W., Wang, X., Qiao, Y., Yan, J., Gao, C., and Sang, N. (2021). Temporal context aggregation network for temporal action proposal refinement. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 485–494, Los Alamitos, CA, USA. IEEE Computer Society.

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

Shi, D., Zhong, Y., Cao, Q., Ma, L., Lit, J., and Tao, D. (2023a). Tridet: Temporal action detection with relative boundary modeling. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18857–18866.

Shi, D., Zhong, Y., Cao, Q., Ma, L., Lit, J., and Tao, D. (2023b). Tridet: Temporal action detection with relative boundary modeling. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18857–18866.

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.

Yang, L., Peng, H., Zhang, D., Fu, J., and Han, J. (2020). Revisiting anchor mechanisms for temporal action localization. *IEEE Transactions on Image Processing*, 29:8535–8548.

Zhang, C.-L., Wu, J., and Li, Y. (2022). Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*, volume 13664 of *LNCS*, pages 492–510.

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: