Multisports challenge 2022 report: Spatio-Temporal Action Detection Under Large Motion

Gurkirt Singh Vasileios Choutas Suman Saha Fisher Yu
Luc Van Gool
Computer Vision Lab, ETH Zürich

Abstract

Current methods for spatio-temporal action tube detection often extend a bounding box proposal at a given keyframe into a 3D temporal cuboid and pool features from nearby frames. However, such pooling fails to accumulate meaningful spatio-temporal features if the position or shape of the actor shows large 2D motion and variability through the frames, due to large camera motion, large actor shape deformation, fast actor action and so on. In this work, we aim to study the performance of cuboid-aware feature aggregation in action detection under large action. Further, we propose to enhance actor feature representation under large motion by tracking actors and performing temporal feature aggregation along the respective tracks. We define the actor motion with intersection-over-union (IoU) between the boxes of action tubes/tracks at various fixed time scales. The action having a large motion would result in lower IoU over time, and slower actions would maintain higher IoU. We find that track-aware feature aggregation consistently achieves a large improvement in action detection performance, especially for actions under large motion compared to cuboid-aware baseline. As a result, we also report state-of-the-art on the large-scale MultiSports dataset.

1. Introduction

Our MultiSports challenge submission is based entirely on Singh[9]. In which, we aim to study cuboid-aware action detection under varying degree of action instance motion using MultiSports [4] dataset which contains instances with large motions unlike AVA [2] as shown in Fig. 1. We are in process of releasing source code at `https://github.com/gurkirt/ActionTrackDetectron`.

2. Methodology

In this section, we provide more details for our proposed method to handle actions with large motions, which we call

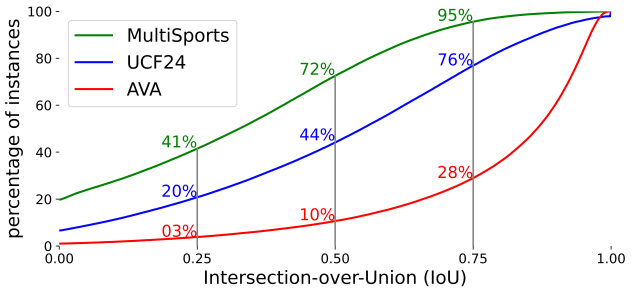

Figure 1: Cumulative density function of IoU measurements for ground-truth bounding box pairs taken one second apart in the training sets of AVA, UCF24, and MultiSports, plotted as percentage of instances falling in cumulative bins shown on the Y-axis. For example, 20% of MultiSports instances has an IoU less than or equal to 0.0 signifying that 20% of instances has very large motion present. In contrast, only 10% of AVA instances has an IoU less than 0.5, meaning that 90% of its instances have a large overlap after one second, *i.e*. large amount instance has small motion.

Track Aware Action Detector (TAAD) [9]. Our method is exactly is the same as TAAD) [9] Here we provide more details on tube construction which are missing in original paper. An overview diagram is shown in Figure 2.

2.1. Tube Construction

Video-level tube detection requires the construction of action tubes from per-frame detections. This process is split into two steps [7], one to link the proposal to form tube hypothesis (i.e. action-tracks), second, these hypothesis need to be trimmed to the part where action is present in these tracks. One can think of these two steps as tracking step plus temporal (start and end time) action detection step. The majority of the existing action tube detection methods [4, 5, 8, 10] use a greedy proposal linking algorithm first proposed by in [3, 11] for the first step. For baseline approach, we use the same method for tube linking process from [11]. Whereas for our method (TAAD), we already

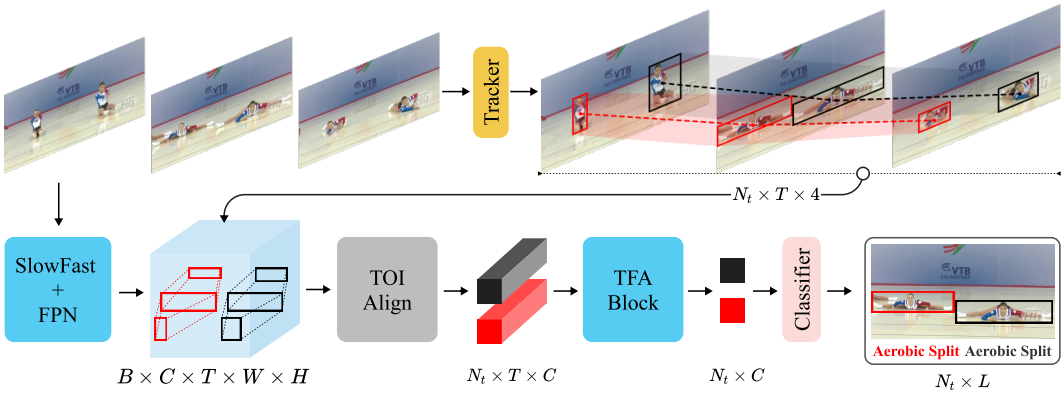

Figure 2: Proposed Track Aware Action Detector (TAAD): Given an input clip with T frames, we extract features using a video recognition network [1] and N_t per-actor tracks from a tracker. The TOI-Align operation extracts per-track features from the entire video sequence, using an RoI-Align operation and the track boxes, returning a $N_t \times T \times C$ feature array. Next, the Temporal Feature Aggregation (TFA) module aggregates the features along the temporal dimension and passes the resulting $N_t \times C$ array to the action classifier that predicts the action label.

have tracks, as a result, the linking step is already complete. While the temporal trimming of action-tracks is performed using label-smoothing optimisation [7], which is used by many previous works [3, 5], specifically, we use class-wise temporal trimming implementation provided by [11].

More specifically, we implement temporal trimming with the help of 'python' implementation of [11] [1]. We score a each box of a track in multiclass setting using TAAD. Then we perform temporal trimming of each track for each class individually. Given a t frames long tack $T = \{(V_1, b_1), ...(V_i, b_i), ...(V_t, b_t)\}$, where b_i is bounding box and V_i is score vector of c classes. For a class c, a binary label $l_i \in \{c, 0\}$, is need to be assigned, then temporal trimming is reduced to finding an optimal binary labelling $L = \{l_1, ...l_i, ...l_t\}$. This can be achieved by solving following optimisation formulation:

$$E(l) = \sum_{i=1}^{i=t} V_i(b_i) - \sum_{i=2}^{i=t} \psi(l_i, l_{i-1}) \qquad (1)$$

where $V_i(b_i) = V_i(c)$ if $l_i = c$, $1 - V_i(c)$ if $l_i = 0$. The pairwise potential ψ is defined as: $\psi(l_i, l_{i-1}) = 0$ if $l_i = l_{i-1}$, $\psi(l_i, l_{i-1}) = \alpha_c$ otherwise. We cross-validated α_c on validation set for each class.

3. Experiments

In this section, we evaluate our TAAD method on MultiSports. Table 1 show results on validation and test set.

[1] https://github.com/gurkirt/3D-RetinaNet

Table 1: Comparison of action detection performance of the proposed method to our baseline model and other state-of-the-art methods on MultiSports dataset. TAAD combined with TFA modules leads to state-of-the-art detection performance. Our TAAD model is trained only on training-set.

Method	f-mAP 0.5	v-mAP 0.2	v-mAP 0.5	v-mAP .1:.9
Validation-Set				
YOWO [4, 5]	25.2	12.9	9.7	–
MOC [4, 5]	25.2	12.9	9.7	–
SlowFast-R50 [1, 4]	27.7	24.2	9.7	–
SlowFast-R101 [6]	29.5	28.1	8.4	12.3
SlowFast-R101+PCCA [6]	42.2	41.0	20.0	20.9
Baseline (ours)	49.6	54.1	31.3	28.9
Baseline + tracks (ours) †	50.6	56.3	33.0	30.9
TAAD + MaxPool (ours)	53.9	58.6	34.8	32.4
TAAD + ASPP (ours)	54.4	59.2	36.0	33.0
TAAD + TCN (ours)	**55.3**	**60.6**	**37.0**	**33.7**
Test-Set				
TAAD + TCN (ours)	**51.6**	**56.4**	**33.8**	**31.7**

* evaluated using tracks at test time.

4. Conclusion

We observe that existing action detection methods struggle in the presence of large motions, *e.g.* motion due to fast actor movement or large camera motion. We introduce Track Aware Action Detector (TAAD), a method that utilizes actor tracks to solve this problem. TAAD aggregates information across actor tracks, rather than using a tube made of cuboid from proposal boxes. TAAD not only bridges the performance gap between motion categories, but

also sets a new state-of-the-art for MultiSports by beating last year's challenge winner by a large margin.

References

[1] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. SlowFast networks for video recognition. In *Computer Vision and Pattern Recognition (CVPR)*, pages 6202–6211, 2019.

[2] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Actions . In *Computer Vision and Pattern Recognition (CVPR)*, pages 6047–6056, 2018.

[3] Vicky Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, and Cordelia Schmid. Action Tubelet Detector for Spatio-Temporal Action Localization. In *International Conference on Computer Vision (ICCV)*, 2017.

[4] Yixuan Li, Lei Chen, Runyu He, Zhenzhi Wang, Gangshan Wu, and Limin Wang. MultiSports: A Multi-Person Video Dataset of Spatio-Temporally Localized Sports Actions. In *International Conference on Computer Vision (ICCV)*, pages 13536–13545, 2021.

[5] Yixuan Li, Zixu Wang, Limin Wang, and Gangshan Wu. Actions as moving points. In *European Conference on Computer Vision (ECCV)*, 2020.

[6] Zhiqing Ning, Qiaokang Xie, Wengang Zhou, Liangwei Wang, and Houqiang Li. Person-Context Cross Attention for Spatio-Temporal Action Detection. Technical report, Huawei Noah's Ark Lab, and University of Science and Technology of China, 2021.

[7] Suman Saha, Gurkirt Singh, Michael Sapienza, Philip HS Torr, and Fabio Cuzzolin. Deep learning for detecting multiple space-time action tubes in videos. In *British Machine Vision Conference (BMVC)*, 2016.

[8] Gurkirt Singh, Stephen Akrigg, Manuele Di Maio, Valentina Fontana, Reza Javanmard Alitappeh, Suman Saha, Kossar Jeddisaravi, Farzad Yousefi, Jacob Culley, Tom Nicholson, et al. Road: The road event awareness dataset for autonomous driving. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 1(01):1–1, feb 5555.

[9] Gurkirt Singh, Vasileios Choutas, Suman Saha, Fisher Yu, and Gool Luc Van. Spatiotemporal action detection under large motion. *arXiv preprint, arXiv:2209.02250*, pages 0–0, 2022.

[10] Gurkirt Singh, Suman Saha, and Fabio Cuzzolin. TraMNet-Transition Matrix Network for Efficient Action Tube Proposals. In *Asian Conference on Computer Vision (ACCV)*, pages 420–437. Springer, 2018.

[11] Gurkirt Singh, Suman Saha, Michael Sapienza, Philip HS Torr, and Fabio Cuzzolin. Online Real-time Multiple Spatiotemporal Action Localisation and Prediction. In *International Conference on Computer Vision (ICCV)*, pages 3637–3646, 2017.