2022 IEEE 5th International Conference on Multimedia Information Processing and Retrieval (MIPR) | 978-1-6654-9548-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/MIPR54900.2022.00074

Generic Action Start Detection

Yuexi Zhang[1,2],Ming Chen[1], Yikang Li[1], Jenhao Hsiao[1], Octavia Camps[2] and Chiuman Ho[1]
[1] OPPO US Research Center, 2479 E Bayshore Rd, RM 110, Palo Alto, CA 94303
[2] College of Engineering, Northeastern University
{yuexi.zhang, ming.chen,yikang.li1,mark, chiuman}@innopeaktech.com, camps@coe.neu.edu

Abstract

*The online detection of action start in video data has witnessed an increase in attention from both academia and industry, for abundant use-cases (e.g., an alert mechanism in videos used for surveillance with an ability to automate the recording of key frames and timestamp). Conventional approaches heavily rely on frame-level annotations and other prior knowledge that can only be applied to limited categories. In this paper, we introduce **G**eneric **A**ction **S**tart **D**etection (GASD): a new task that aims to detect the taxonomy-free action start in an online manner. Furthermore, one novel yet simple design, 3D MLP-mixer based architecture with a multiscaled sampling training strategy, is proposed, which makes the GASD algorithm favorable for edge-device deployment. The GASD task is validated on two large-scale datasets, THUMOS'14 and ActivityNet1.2. Results demonstrate that the proposed architecture achieves the SOTA performance on the GASD task compared with other online action start detection algorithms.*

1. Introduction

Online action start detection in untrimmed videos is to determine when the specific action starts with minimal latency. This task is important for real applications such as emergency alerts in surveillance systems. Recent years have seen significant progress in temporal action localization (TAL), online action detection (OAD), and online detection of action start (ODAS) in videos. Despite this, the modeling of intrinsic taxonomy-free action characteristics that we humans can naturally perceive and decide the action starts is still lacking. Existing methods are either requiring information from the entire video or heavily rely on taxonomy-based prior knowledge and dense action classification labels.

Therefore, we propose a new task named the **G**eneric **A**ction **S**tart **D**etection (GASD), which aims to detect the action start point within streaming videos by learning the intrinsic action characteristics rather than the specific pre-defined taxonomy-based information. There are three main contributions in our paper: **(1)** We propose a new task GASD for overcoming existing issues in both related online and offline action start detection algorithms; **(2)**We design a edge device friendly 3D MLP-mixer based method for the GASD task; **(3)**The proposed method outperforms other online action start algorithms in two benchmark datasets and the GASD has more potentials in real-world applications.

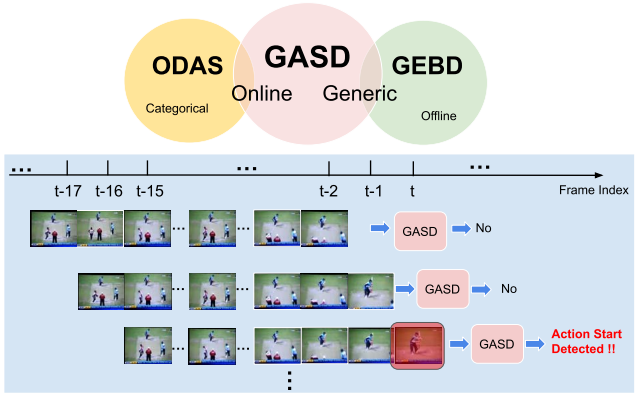

Figure 1: **Illustration of GASD and its relationship with ODAS, GEBD.** ODAS focuses on localizing the start point of specific action categories in untrimmed, streaming video; GEBD aims to detect generic, taxonomy-free event boundaries that segment a whole video into chunks and requires all the information from the full video and can only be performed offline. The newly proposed GASD aims for online action start detection on generic scenarios without the limitation of action categories.

2 Related Work

TAL task aims to localize all the temporal boundaries of interested actions in untrimmed videos. Full information from the videos is required that can only be conducted under offline settings. Dense temporal annotations for every action instance is also needed for better performance [2, 3].

ODAS is to distinguish the specific start of action of

2770-4319/22/$31.00 ©2022 IEEE
DOI 10.1109/MIPR54900.2022.00074

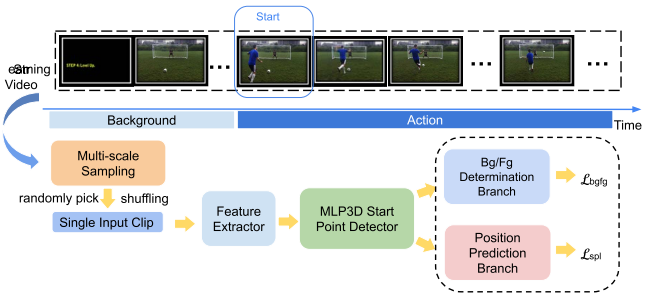

Figure 2: **Overview of the whole joint-training architecture** video clips are first sampled by multiscale sampling strategy, I3D features are extracted for each clip and input into the 3D MLP-mixer based start point detection model, two loss branches, the BG/FG branch, and the SPL branch jointly determine the if there exists an action start and predicts the index within the clip.

interests from its proceeding complex background. Solutions for the ODAS task [4, 5, 10] usually combine a frame-level action classification with a start point localization. Extensive frame-level labeling (including action category and start point annotation) is required.

Generic Event Boundary Detection (GEBD) [9] is newly proposed that aims to detect generic, taxonomy-free event boundaries that segment a whole video into chunks. It is designed to localize the moments where humans naturally perceive event boundaries and scales to generic videos. Similar to the TAL problem, as a boundary detection task, GEBD also requires all the information from the full video and can only be performed offline.

3 Methodology

Given an untrimmed, streaming video, the GASD task is to detect the action start as soon as it occurs in an online manner without prior knowledge of action categories. We proposed a novel yet simple design(see Figure. 2), a 3D MLP-mixer based architecture with a multi-scaled sampling strategy, which is favorable for the edge deployment.

3.1 Multi-scale Sampling and Feature Extraction

Given facts that (1) datasets include action instances with starting points annotations are scarce, (2) action instances have different speeds and duration, and (3) the difference in appearance and feature space between adjacent video clips is trivial, we adopt multiscale sampling scheme for data augmentation as well as multiscale temporal information modeling. Video clips are sampled at various durations and fixed t frames are uniformly sub-sampled from each video clip. Sub-sampled video clips are either labeled as 'foreground(FG)' which contains the action start point or 'background(BG)' otherwise. Those "BG/FG" labels will be utilized in the training stage. The I3D network [1] pre-trained on Kinetics is used as a feature extractor. Each sub-sampled video clip is fed into the I3D feature extractor to obtain a 3D appearance feature representation which is then taken as the input for the 3D MLP-mixer architecture.

3.2 3D MLP-mixer based Architecture

An MLP-mixer based architecture, MLP3D, is proposed for the GASD task. The MLP3D can be regarded as an extension of the original MLP-mixer architecture [11] into the 3D dimension which models the temporal relationship. Different from the original MLP-mixer, which does a channel fusion process immediately after the image token mixing layer, the proposed MLP3D indeed takes another temporal token mixing process between the image token mixing and the channel fusion process. The 3D mixer layer can be described as:

$$\begin{aligned} H_{*,*,i} &= X_{*,*,i} + \mathbf{MLP}_1(X_{*,*,i}), i \in [1, C], \\ U_{*,j,*} &= H_{*,j,*} + \mathbf{MLP}_2(H_{*,j,*}), j \in [1, T], \\ Y_{k,*,*} &= U_{k,*,*} + \mathbf{MLP}_3(U_{k,*,*}), k \in [1, S], \end{aligned} \tag{1}$$

where X is the input of the MLP3D mixer layer, which is a 4-dimension features (batch size, temporal axis, frame token axis, embedding channel axis), C is the number of image patch tokens for each frame, T is the temporal length of the input feature, and S is the dimension of the embedding channels. The MLP_1, MLP_2, and MLP_3 is the MLP structure for frame token axis, temporal axis and the channel axis respectively. The MLP structure is consists of two fully-connected layers connecting by a GELU nonlinearity activation layer.

There are two outputs from the proposed MLP3D architecture: BG/FG classification and the start point localization(SPL). The SPL output returns the action start point in frame-level once the current input clip triggers the FG determination.

To coordinate two outputs of the proposed architecture, two kinds of focal losses [8] are utilized due to imbalanced data and hard negative samples in the GASD task. For the BG/FG output branch, a binary focal loss is implemented as $\mathcal{L}_{bgfg} = -\alpha(1-P)^{\gamma}\mathbf{Y}log(P) - (1-\alpha)P^{\gamma}(1-\mathbf{Y})log(1-P)$,where α is the weight for the imbalanced data distribution of background and foreground sample clips, γ is the hyper-parameter for hard negative samples, $\mathbf{Y}$ is the ground truth label for the foreground, and P is the probability of predicting the positives. For the SPL branch on the start frame localization, it can be treated as a classification problem with (t+1) classes, where t is the frame number together with a background class. The multi-class focal loss is utilized: $\mathcal{L}_{spl} = -\alpha \sum_{\tau=1}^{t+1} \tau_{P'}(1 - P'_{\tau})^{\gamma} log(P'_{\tau})$, where $\tau_{P'}$ is the one-hot ground truth label for denoting the action start frame and P'_{τ} is the probability of predicting the action start.

Therefore, the total loss for the entire framework is computed as: $\mathcal{L} = \mathcal{L}_{spl} + \lambda * \mathcal{L}_{bgfg}$. Where λ is the trade-off hyper-parameter between the BG/FG outputs and the SPL outputs during the joint training.

4 Experiments

In this section, we presented the evaluation protocols and experimental results of our proposed model on two benchmark datasets: **THUMOS'14 [7]** and **ActivityNet v1.2 [6]**. THUMOS'14 contains 20 sports classes and each action instance is temporally annotated with start/end timestamp. Following the previous work [4, 5, 10], we trained our model on the validation set with 200 untrimmed videos including 3K action instances, and evaluated on the test set with 214 untrimmed videos including 3.3Ks action instances. ActivityNet v1.2 involves 100 actions with an average of 1.5 action instances per video. We trained our model on the validation set with 4819 untrimmed videos and evaluated on the test set with 2383 untrimmed videos.

4.1 Implementation Details

As mentioned in section 3, each sub-sampled video clip is fixed to $t = 16$ frames. For untrimmed videos, we conduct temporal sliding windows at varied durations (e.g., [16, 32, 64, 128, 256, 512] frames) with 75% overlap. The BG/FG sample ratio is set to $3:1$. The multiscale sampling strategy is disabled during the inference stage that each testing video clip is sampled at consecutive frames. Given that the frame index prediction might need the temporal order information within the clip, the parameters within the I3D network are frozen from the first layer to the intermediate i3d.mixed4f layer, which leads to an output with a shape of $4 \times 832 \times 14 \times 14$. α is set to be 0.5 and 0.25 for $\mathcal{L}_{bgfg}$ and $\mathcal{L}_{spl}$ respectively, γ is set to 2 in both focal loss equations. λ in the loss function $\mathcal{L}$ is fixed at 1 for the best reported model. Batch size is set to be 48. We use SGD as optimizer and set the initial learning rate to be 1×10^{-2}. For the MLP3D model, we set the depth of the mixer layer as 6. The method is implemented with pytorch.

4.2 Evaluation Protocol

The outputs from the BG/FG branch and the SPL branch jointly determine the start point during the inference stage, the final start point is only generated if the input clip is predicted to be a foreground, and the predicted start point index is within range.

We use precision/recall as offline evaluation metric to test the performance on the FG/BG branch as its performance determines the coarse-grained precision on the start point detection. For THUMOS'14 the precision is **67.5%**, the recall is **48%** and for ActivityNet 1.2 the precision is **59.9%**, the recall is **17.0%**. The low recall on ActivityNet is due to more complicated background in videos.

The point-based average precision (P-AP) [10], is used as the online evaluation metric. The P-AP evaluates the precision of correct action start predictions at different temporal tolerance in an online fashion. Similar to segment-level average precision, no duplicate detections are allowed for the same ground-truth point. During the testing stage, we evaluated the video clips by their temporal order, measured the average precision (AP) in each video under specific temporal tolerance, and the final P-AP is obtained by the average of all the testing videos' results.

The online test results are summarized in Table 1. The SOTA method on ODAS task, WOAD [1] , is used as a comparison. The results from WOAD on THUMOS'14 are obtained by using their official codes and models but removing the action classification part.As demonstrated in Table 1, our methods outperformed the WOAD method on all the temporal offsets. In addition, for ActivityNet1.2, there are no available official codes for ODAS task so we only report our GASD model results.

4.3 Ablation Study

Our performance improvements benefit from (1) the multiscale sampling strategy (2) the joint training of BG/FG and SPL branch (3) the novel MLP-based start point detection model.The contribution of each module/strategy in the proposed solution is summarized in Table 2. To make the results more consistent, the offline tests on THUMOS'14 are used to evaluate all the model variants.The joint training of SPL and BG/FG branch will improve the performance of GASD by $3\%-4\%$ due to the additional regularization. The new proposed MLP3D algorithm together with the multiscale sampling scheme will further gain 4% improvement in precision as well.

We also compared our proposed method with WOAD on efficiency. Though our proposed methods has more parameters (136M) compared to WOAD (110 M), our FLOPS is 1.6G, which is only **half** of the WOAD model's. Moreover, our MLP-based method is more easily to be optimized (e.g., weight pruning or integer aware training), making it a more edge device friendly algorithm. Take another closer look at Table 1, it also shows the limitation of directly applying the current ODAS methods to the newly proposed GASD tasks. The ODAS related solutions strongly depend on the category-dependent prior knowledge and frame-level annotations, which is unavailable for GASD task. Therefore, the start point localization module from the ODAS task is not as capable as our proposed method for the GASD task.

Moreover, our proposed GASD algorithm is versatile and can be easily extended for the ODAS task. Either a Multi-class action classification branch or a third loss to learn the action categories can be added on the top of the current GASD architecture. Adopting the same evaluation

[1] https://github.com/salesforce/woad-pytorch

Evaluation on THUMOS'14 Dataset														
Offsets (second)	0.1	0.3	0.5	0.8	1	2	3	4	5	6	7	8	9	10
WOAD	0.9	13.1	14.0	14.1	14.1	14.1	14.1	14.1	14.1	14.1	14.1	14.1	14.1	14.1
Ours	**4.1**	**13.3**	**14.1**	**16.1**	**17.9**	**21.2**	**22.4**	**25.7**	**26.2**	**26.7**	**28.8**	**29.1**	**29.3**	**29.5**
Evaluation on ActivityNet 1.2														
Ours	-	-	-	-	7.4	16.9	17.5	17.7	18.1	18.1	18.1	18.1	18.1	18.1

Table 1: **Evaluation on THUMOS'14, ActivityNet 1.2** For THUMOS'14, the results from WOAD on GASD task is obtained by using their pretrained models with official codes and removing the action classification part. For both two datasets, using p-mAP at depth Rec=1.0.

criteria, our modified GASD model can achieve competitive performance with **19.6%-45.5%** mean P-AP over 1-10s offset. The reported SOTA performance on ODAS tasks are [4] **19.5%-39.8%** mean P-AP with appearance feature only and [5] **21.9%-53.1%** mean P-AP with two-stream features over 1-10s offset on THUMOS'14 dataset. The performance gap between our modified GASD and WOAD [5] on the ODAS task owes to the fact that only the appearance feature is used in our GASD model while additional motion features will also contribute to the ODAS task.

5 Conclusion

In this paper, we propose a new task GASD that focuses on learning the intrinsic action characteristics for taxonomy-free action start detection. Our proposed architecture, MLP3D, is designed to independently determine the action start point for each input video clip and thus has low demand for prior knowledge from previous frames. The experiment results demonstrate that the proposed method achieves the State-of-the-Art performance on the GASD task with high efficiency that enables the potential for edge device deployment.

References

[1] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6299–6308, 2017.

[2] Y.-W. Chao, S. Vijayanarasimhan, B. Seybold, D. A. Ross, J. Deng, and R. Sukthankar. Rethinking the faster r-cnn architecture for temporal action localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.

[3] X. Dai, B. Singh, G. Zhang, L. S. Davis, and Y. Q. Chen. Temporal context network for activity localization in videos. CoRR, abs/1708.02349, 2017.

[4] M. Gao, M. Xu, L. S. Davis, R. Socher, and C. Xiong. Startnet: Online detection of action start in untrimmed videos. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October 2019.

[5] M. Gao, Y. Zhou, R. Xu, R. Socher, and C. Xiong. Woad: Weakly supervised online action detection in untrimmed videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 1915–1923, June 2021.

[6] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 961–970, 2015.

[7] H. Idrees, A. R. Zamir, Y. Jiang, A. Gorban, I. Laptev, R. Sukthankar, and M. Shah. The THUMOS challenge on action recognition for videos "in the wild". CoRR, abs/1604.06182, 2016.

[8] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision, pages 2980–2988, 2017.

[9] M. Z. Shou, D. Ghadiyaram, W. Wang, and M. Feiszli. Generic event boundary detection: A benchmark for event segmentation. CoRR, abs/2101.10511, 2021.

[10] Z. Shou, J. Pan, J. Chan, K. Miyazawa, H. Mansour, A. Vetro, X. Giró-i-Nieto, and S. Chang. Online action detection in untrimmed, streaming videos - modeling and evaluation. CoRR, abs/1802.06822, 2018.

[11] I. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. arXiv preprint arXiv:2105.01601, 2021.

module	SPL	FGBG	multiscale	Precision(%)
3D CNN	✓			59.5
3D CNN	✓	✓		63.7
3D CNN	✓	✓	✓	64.5
MLP3D	✓	✓	✓	67.5

Table 2: **Ablation study of joint training** of SPL and BG/FG branch with different backbone models.