Refining Action Segmentation with Hierarchical Video Representations -Supplementary Material-

Hyemin Ahn[1] and Dongheui Lee[1,2]
[1]German Aerospace Center (DLR) [2]Technical University of Munich
{hyemin.ahn, dongheui.lee}@dlr.de

1. Experiments

In this supplementary material, we show additional qualitative results that could not be shown on the original manuscript due to the page limit. In addition, we compare our method with Graph-based Temporal Reasoning Module (GTRM) [4], which is another refiner that can also improve the performance of action segmentation backbone models. Also, we show additional quantitative results related to the Section 4.3.2 in our original manuscript, which use SSTDA [1] or ASRF [5] as unseen backbone models for test while using other models for training our HASR. For example, when ASRF is the unseen model for test, we use segmentation results from MS-TCN [3], SSTDA, and GRU-based model to train our HASR.

1.1. Additional Qualitative Results

Figure 1 shows the additional example refinement results from the proposed Hierarchical Action Segmentation Refiner (HASR). Figure 1(a) shows how HASR refines the segmentation results from ASRF. The given video is from the Breakfast dataset [6], and it is about a human making fried eggs. Even if the frame-level feature information represents the human cooking eggs, the result shows that the action segmentation backbone model (ASRF) misunderstands 'egg' as 'dough' or 'pancake'. The refinement result shows that our proposed HASR is able to correct the segment action labels which do not match the frame-level feature information. We suppose this is due to our segment-level representation, which can encode the frame-level features consisting the action segment. We believe that segment-level representations enabled our HASR to reinterpret the given segments, and correct the wrong segment results that do not match the frame-level features.

Figure 1(b) shows another result when our HASR refines the result from the SSTDA. The given video is from the 50Salads dataset [8], and it shows an egocentric video from a human when making a salad. In this video, the human cuts tomato, cucumber, lettuce, and cheese in order. The action segmentation backbone model (SSTDA) predicts that the human cuts lettuce for a while. But it estimates that the human would suddenly peel cucumber, place cucumber into bowl, then place tomato into bowl, even if the cucumber and tomato were already mixed into the salad bowl. The refinement result shows that our HASR successfully corrects these false segmentation results, which are unnatural action sequences for making a salad.

Figure 1(c) shows the result when our HASR refines the result from MS-TCN, which is the unseen backbone model. In other words, in this experiment, the HASR was trained to refine the action segmentation results from ASRF, SSTDA, and GRU-based models, and employed to refine the action segmentation results from MS-TCN. Here, the input video is about preparing a tea, from Breakfast datset. The result shows that the action segmentation backbone model (MS-TCN) completely misunderstands the whole video. But our HASR successfully corrects these wrong segmentation results, and we claim it is also due to our segment-level representations, which reinterprets the frame-level features consisting the segments.

However, as shown in Figure 1(d), our HASR can also fail even the action segmentation backbone model predicts the correct segmentation results. Here, the input video is about preparing a coffee, from Breakfast dataset, and the ASRF was used as the action segmentation backbone model. The result shows that our HASR misunderstands the action of 'spoon sugar' as 'pour sugar', which shows that our segment-level representations do not always interpret the video correctly. But still, we claim that our HASR can be an effective tool for improving the performance of action segmentation models. Based on the significant performance gains that were reported in our original manuscript, it can be understood that our HASR succeeds more often than makes this kind of mistake.

1.2. Comparison with Graph-based Temporal Reasoning Module (GTRM) [4]

In this section, we compare the performance gain between our HASR and Graph-based Temporal Reasoning Module (GTRM) [4]. GTRM is another refiner based on the

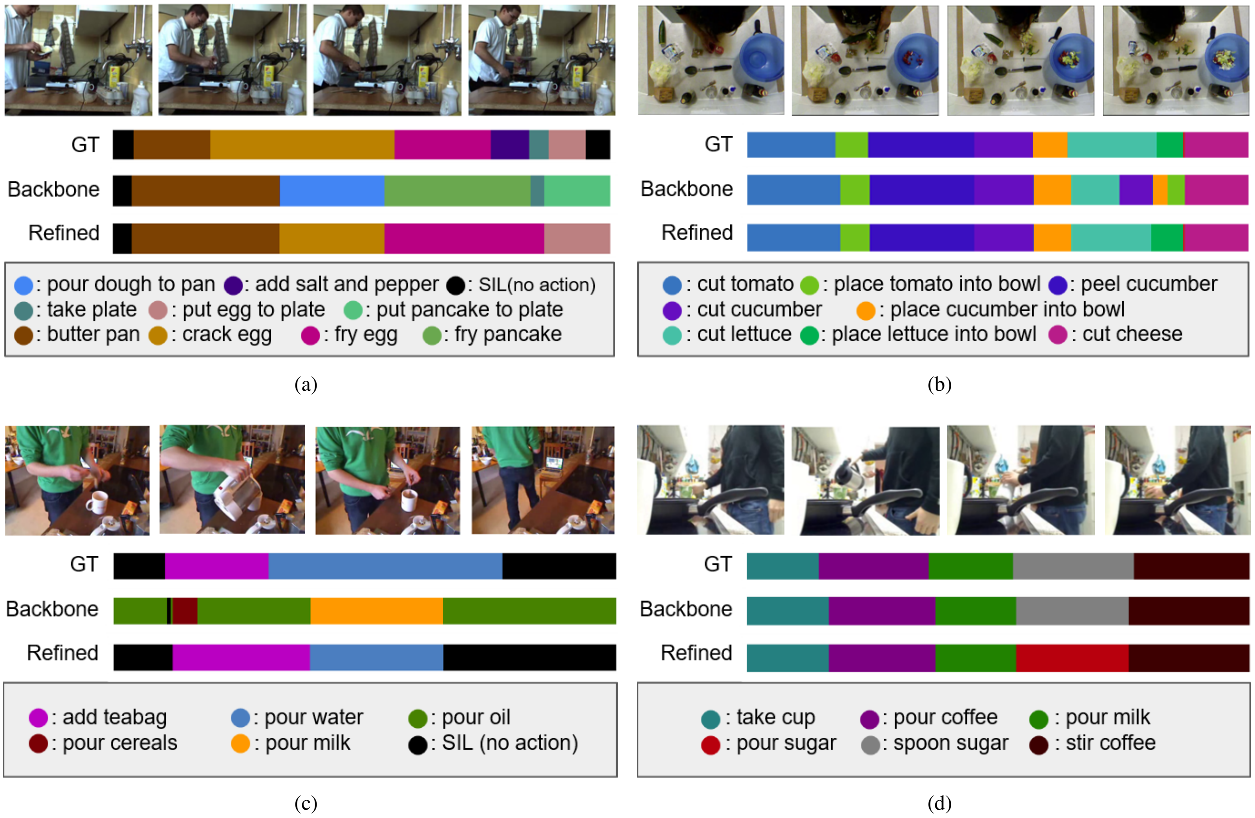

Figure 1. Additional qualitative results from HASR with various backbone models and datasets. Best view in color. (a) Refinement from ASRF with Breakfast dataset. (b) Refinement from SSTDA from 50Salads dataset. (c) Refinement from MS-TCN from Breakfast dataset, when MS-TCN is the unseen action segmentation backbone model. (d) Failure case from ASRF with Breakfast dataset.

	50Salads					**Breakfast**				
	F1@{0, 25, 50}			Edit	Acc	F1@{0, 25, 50}			Edit	Acc
MS-TCN (impl. from [4])	73.4	71.0	61.5	67.2	80.2	57.3	53.4	41.4	58.8	60.0
MS-TCN + GTRM [4]	75.4	72.8	63.9	67.5	82.6	57.5	54.0	43.3	58.7	65.0
Gain	2.0	1.8	1.4	0.3	**2.4**	0.2	0.6	1.9	-0.1	**5.0**
MS-TCN (our impl.)	77.2	74.7	64.8	70.4	80.3	63.5	58.3	45.9	66.2	67.7
MS-TCN + HASR	83.4	81.8	71.9	77.4	81.7	73.2	67.9	54.4	70.8	69.8
Gain	**6.2**	**7.1**	**7.1**	**7.0**	1.4	**9.7**	**9.6**	**8.6**	**4.6**	2.0

Table 1. Performance gain comparison between HASR and GTRM [4] based on 50Salads and Breakfast datset. The performance records of GTRM are from their original paper [4].

graph convolutional network, which can improve the performance of the action segmentation backbone models such as MS-TCN [3]. It refines the segmentation results from the action segmentation backbone model considering the relation of each action segment with its neighboring segments.

In the paper of GTRM [4], the authors mainly conducted their experiments based on the EGTEA [7] and EPIC-KITCHEN [2] dataset. In addition, they conducted the experiments based on the 50Salads and Breakfast dataset, when the action segmentation backbone model is MS-TCN [3]. Therefore, as shown in Table 1, we compare their official performance records with ours, based on 50Salads and Breakfast datasets when the backbone model is MS-TCN. Note that the performances of the backbone model (MS-TCN) are different since we and [4] trained the model individually.

The results show that the performance gains of GTRM is higher than ours with respect to the frame-wise accuracy. This shows that the refinement process by considering the relation between segments is meaningful as [4] suggested. However, when considering the segmental edit and F1 scores, the performance gain of our HASR is generally

GTEA					
Method	F1@{10, 25, 50}			Edit	Acc
SSTDA	91.1	88.8	75.6	87.9	**79.4**
SSTDA+HASR	**91.6**	**89.5**	**77.2**	**88.4**	**79.4**
ASRF	87.9	86.1	**75.2**	81.9	**77.1**
ASRF+HASR	**88.4**	**86.6**	74.2	**82.2**	76.9
50Salads					
Method	F1@{10, 25, 50}			Edit	Acc
SSTDA	80.6	78.7	70.8	74.9	**82.5**
SSTDA+HASR	**83.6**	**82.2**	**74.5**	**77.7**	82.4
ASRF	85.1	83.3	**77.7**	**79.9**	**83.7**
ASRF+HASR	**85.6**	**84.4**	76.8	79.2	83.5

Table 2. Refinement results when SSTDA or ASRF are used as an unseen backbone models for test.

higher than that of GTRM, even if the performance of the backbone model (MS-TCN) is higher from our implementation. This shows that our HASR solves problems such as over-segmentation better than GTRM, and the quality of the segments refined by HASR is higher. Based on this result, we would like to highlight that solving the action segmentation refinement problem can be more effective if the refiner can consider the hierarchical video representations as we suggest.

1.3. Additional Quantitative Results for Sec. 4.3.2

Table 1.3 shows the additional refinement results when SSTDA or ASRF are used as unseen backbone models for test while others are used for training our HASR. For example, when using ASRF as the unseen backbone model for test, we use segmentation results from MS-TCN, SSTDA, and GRU-based model for training our HASR. The results show that the performance gain of unseen SSTDA is larger than the one when SSTDA was used for training HASR (Check Table 1 and 2 in our original manuscript). However, it is shown that the performance gain of unseen ASRF is lower than our expectation, which implies that the dataset collected from ASRF is critical when training our HASR with ASRF as a backbone segmentation model. But still, we would like to claim that the performances are improved in general, which implies that HASR can correct out-of-context segment labels for unseen backbone models such as SSTDA and ASRF, which is more advanced compared to MS-TCN and GRU-based model.

Note that these experiment results are based on the GTEA and 50Salads dataset. We were not able to obtain the result from Breakfast dataset since it is very time consuming. The relevant results will be released to our code repository webpage[1].

[1] https://github.com/cotton-ahn/HASR_iccv2021

References

[1] Min-Hung Chen, Baopu Li, Yingze Bao, Ghassan AlRegib, and Zsolt Kira. Action segmentation with joint self-supervised temporal domain adaptation. In *Proceedings of the conference on computer vision and pattern recognition*, pages 9454–9463, 2020. 1

[2] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision*, pages 720–736, 2018. 2

[3] Yazan Abu Farha and Jurgen Gall. MS-TCN: Multi-stage temporal convolutional network for action segmentation. In *Proceedings of the conference on computer vision and pattern recognition*, pages 3575–3584, 2019. 1, 2

[4] Yifei Huang, Yusuke Sugano, and Yoichi Sato. Improving action segmentation via graph-based temporal reasoning. In *Proceedings of the conference on computer vision and pattern recognition*, pages 14024–14034, 2020. 1, 2

[5] Yuchi Ishikawa, Seito Kasai, Yoshimitsu Aoki, and Hirokatsu Kataoka. Alleviating over-segmentation errors by detecting action boundaries. In *Proceedings of the winter conference on applications of computer vision*, pages 2322–2331, 2021. 1

[6] H. Kuehne, A. B. Arslan, and T. Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the conference on computer vision and pattern recognition*, 2014. 1

[7] Yin Li, Miao Liu, and James M Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of the European Conference on Computer Vision*, pages 619–635, 2018. 2

[8] Sebastian Stein and Stephen J McKenna. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 729–738, 2013. 1