Supplementary Material: Weakly-Supervised Online Action Segmentation in Multi-View Instructional Videos

Reza Ghoddoosian[1,2*], Isht Dwivedi[1], Nakul Agarwal[1], Chiho Choi[1], and Behzad Dariush[1]
[1]Honda Research Institute, USA
[2]VLM Lab, University of Texas at Arlington
reza.ghoddoosian@mavs.uta.edu, {idwivedi,nagarwal,cchoi,bdariush}@honda-ri.com

1. Overview

In this supplementary material we first provide the definitions of all the terms used in the paper, explain complexity as a limitation, and then show and discuss more qualitative segmentation results of different methods.

2. Glossary of Symbols

We provide specific definitions of symbols in Table 1 for readers to refer to.

3. Limitation

Here we discuss the practical run-time and computational complexity of our method, both in test and training, and compare it with CDFL [4] as an offline baseline. The lower frame rate compared to a greedy approach and the lengthy training-time are notable limitations of the proposed online segmentation method. We hope further work can mitigate this limitation.

3.1. Runtime Frame Rate Analysis

Both our method and the greedy approach [2] compute optical flow (OF) and use the I3D network to extract features. We extracted features on 320×240 frames of the *BD* dataset recorded at 15 fps. On a single GeForce GTX 1080, the OF and I3D network process videos at 90 and 20 fps, respectively. Practically, if online inference is done every 15 frames, then our method segments videos at 10+ fps. While this is less than the 100 fps of the greedy alg., it leads to considerably more accurate results

3.2. Computation Complexity of Online vs. Offline

The complexity of the proposed online inference to fully segment a video of length T and maximum transcript length of N is the same as the offline inference of CDFL ($O(T^2N)$). In other words, online inference at each time step takes $O(TN)$. This is due to DP as the inference at time t depends on the optimal results of previous time steps which have already been obtained as part of DP.

With this in mind, the training complexity of our method over K classes is the sum of complexities for the offline inference ($O(T^2N)$), baseline offline segmentation loss $\mathcal{L}_b$ ($O(\Delta^2 NK)$) and $\mathcal{L}_{OODL}$. $\Delta \ll T$ is a small window size of 10 [4]. A naive implementation of OODL has complexity of $O(T^2N)$. However, if the online and offline inferences are done together outside Alg.1 and $\mathcal{E}(t)$ is summed over segments rather than frames, the OODL complexity becomes $O(TN)$. Hence, regardless of the implementation choice, our overall training has the same complexity as CDFL ($O(T^2N + \Delta^2 NK)$). Our time complexity during test time with M training transcripts is $O(T^2NM)$ which is also the same as that of CDFL. In order to quantitatively support our calculations, we tested both methods on the 4th split of the *BD* dataset. Our method and CDFL took 26 and 21 hrs to train, respectively. Meanwhile at test time, ours and CDFL took 2.7 and 2.4 hrs to run, respectively.

4. Qualitative Results

In Figure 1, we present two segmentation examples on the IKEA [1] (top) and Breakfast [3] (bottom) datasets. We demonstrate how training using multiple view points has let to more robust segmentation results against full occlusion (top) and extremely bad lighting (bottom). Specifically, the top figure depicts a task where the subject assembles a "side table". This assembly consists of four instances of "spinning leg", where the last instance is fully occluded by the subject's body. The baseline method DP_{on} , that is trained on a single viewpoint, misses most of the action, while training on multi-view correspondence and the OODL loss has enabled our final model ($\text{DP}_{\text{on}} + M + \mathcal{L}_{\text{OODL}}$) to capture nearly the full segment.

In the second example, the dark lighting makes it even hard for a human observer to recognize the ongoing action. Our final method is able to identify the action of "adding

*Work done during Reza's internship at Honda Research Institute, USA

Table 1. Definitions of symbols used in the paper.

Symbol	Definition
$\mathbb{A}$	The set of all actions in the dataset
a_n	Action variable at segment n
a_t	Action variable at frame t
$\hat{a}_t$	Predicted action at time t in an online way
$(\overline{\boldsymbol{a}}_1^N, \overline{\boldsymbol{l}}_1^N)$	Offline inference result
$(\tilde{\boldsymbol{a}}_1^{n(t)}, \tilde{\boldsymbol{l}}_1^{n(t)})$	Online inference result until time t
${}_ic_t$	View confidence weight of video i (anchor) at time t
e_n	Energy score of segment n
$\mathcal{E}_{\pi+}$	Energy score of the valid path π^+
$\mathcal{E}_{\pi-}$	Energy score of the invalid path π^-
$\mathcal{E}_{\text{off}}(t)$	Energy score of the offline or valid path until time t
$\mathcal{E}_{\text{on}}(t)$	Energy score of the online path until time t
$\bar{\mathcal{E}}$	Weighted energy score of a path
F_1	Input feature dimension
F_2	Embedding dimension
K	Total number of videos in the data set
l_n	Duration variable of action a_n
$\mathcal{L}_f$	Final loss
$\mathcal{L}_b$	Baseline offline segmentation loss
$\mathcal{L}_{vc}$	View confidence loss to train WPI
M	Number of action labels in the transcript
N	Number of predicted segments in the video
$p_{\text{on}}()$	Causal probability
$p_{\text{off}}()$	Non-causal probability
$\mathbb{P}^-$	The set of all invalid paths
T	Total number of frames in the video
$\boldsymbol{\tau}$	Video transcript
v_i	Video i
$\mathbb{V}_i$	View adjacency set of video i
V	$K \times K$ view adjacency matrix
ω	Feature window size for Φ_f
$\mathbf{x}_1^T$	Sequence of T frame features
${}_i\mathbf{x}_1^T$	Features of video i
$\eta(n)$	Mapping function from segment to frame number
$\Gamma()$	Half Poisson function
λ_a	Estimated mean length of action a
Φ_c	Compare function
Φ_f	Feature embedding function
θ_c	Parameters of Φ_f and Φ_c
θ_a	Parameters of the action classifier, i.e. GRU
π^+	Valid path or offline segmentation action sequence
π^-	Invalid path

tea bag", where both the offline method DP_{off} and online baselines DP_{on} fail. This is an interesting case, where our model is able to outperform even the offline method. One reason is the flexibility of the proposed online segmentation model in switching between different transcripts in a series of online inferences across different time steps. This allows the predicted sequence of actions to potentially come from a transcript not observed at training time. In contrast, in offline segmentation the sequence of inferred action labels is limited only to the training transcripts.

References

[1] Yizhak Ben-Shabat, Xin Yu, Fatemeh Saleh, Dylan Campbell, Cristian Rodriguez-Opazo, Hongdong Li, and Stephen Gould. The ikea asm dataset: Understanding people assembling furniture through actions, objects and pose. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 847–859, 2021. 1

[2] Mingfei Gao, Yingbo Zhou, Ran Xu, Richard Socher, and Caiming Xiong. Woad: Weakly supervised online action detection in untrimmed videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1915–1923, 2021. 1

[3] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 780–787, 2014. 1

[4] Jun Li, Peng Lei, and Sinisa Todorovic. Weakly supervised energy-based learning for action segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6243–6251, 2019. 1

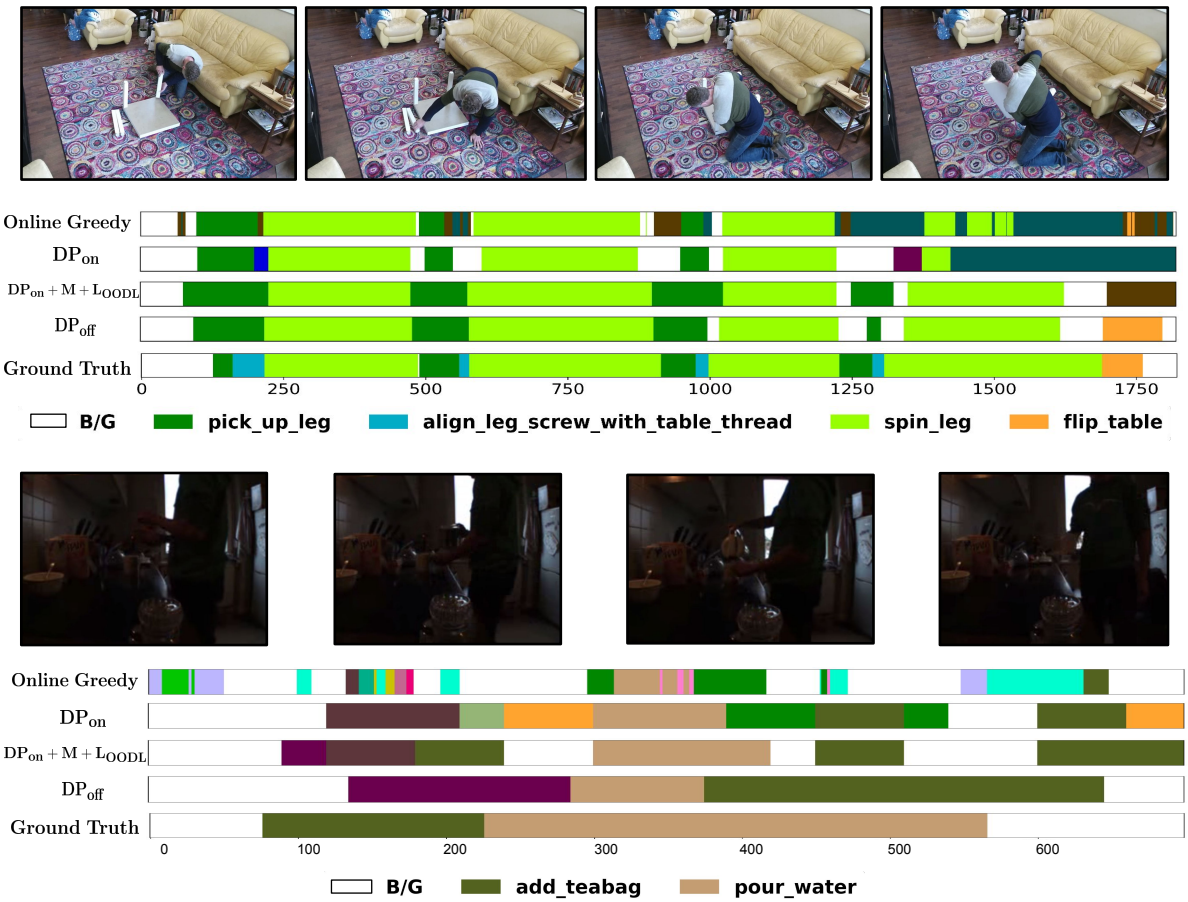

Figure 1. This figure shows segmentation results of various methods on the IKEA (top) and Breakfast (bottom) datasets. Subjects in the top and bottom figures assemble a side table and prepare tea respectively. Legend is shown only for the ground-truth classes.