Temporal Action Segmentation from Timestamp Supervision (Supplementary Material)

Zhe Li, Yazan Abu Farha, Juergen Gall
University of Bonn, Germany

We evaluate our model on additional settings for generating the timestamps annotations. We further analyze the impact of noise on the performance.

A. Frame Selection for the Timestamp Annotations

In the paper, we randomly select one frame from each action segment in the training videos. Table A shows results for two additional settings. While using the center frame of each segment achieves comparable results to a random frame, the performance drops when the start frame is used. Humans, however, would not annotate the start frame since it is more ambiguous (see Fig. 4 in [1]).

Timestamps	F1@{10, 25, 50}			Edit	Acc
Start frame	65.5	52.2	28.0	70.4	51.2
Center frame	**70.8**	63.5	45.4	**71.3**	61.3
Random	70.5	**63.6**	**47.4**	69.9	**64.1**

Table A. Using start, center, or random frame as timestamp for each action on the Breakfast dataset.

B. Human Annotations vs Generated Annotations

In Table B, we directly compare the human annotations from [1] with simulated annotations on the GTEA dataset. The results show that the performance with random sampling is very close to real annotations.

mAP@IoU	0.1	0.3	0.5	0.7	Avg
Human annotations	**60.2**	44.7	**28.8**	**12.2**	**36.4**
Random sampling	59.7	**46.3**	26.0	10.4	35.6

Table B. Human vs. generated annotations on the GTEA dataset.

C. Impact of Noise

In Table 10 in the paper, the timestamps are randomly sampled from the videos. Thus, there are sometimes multiple timestamps for one action and not all actions are annotated. Table C shows the percentage of action segments with 0, 1, or >1 timestamps (TS). As shown in the table, our approach is robust to annotation errors.

Fraction	% actions with 0 TS	% with 1 TS	% with >1 TS	Acc
0.1	3.5	3.4	93.1	68.4
0.01	27.5	21.6	50.9	67.4
0.0032	49.0	25.2	25.8	61.7
0.0032	0	100	0	64.1

Table C. Percentage of actions with 0, 1, or >1 timestamps (TS) using the protocol of Table 10. The last row is the protocol of Table 7.

References

[1] Fan Ma, Linchao Zhu, Yi Yang, Shengxin Zha, Gourab Kundu, Matt Feiszli, and Zheng Shou. SF-Net: Single-frame supervision for temporal action localization. In *European Conference on Computer Vision (ECCV)*, 2020. 1