Supplementary Material

Reza Ghoddoosian Saif Sayed Vassilis Athitsos
Vision-Learning-Mining Lab, University of Texas at Arlington
{reza.ghoddoosian, saififtekar.sayed}@mavs.uta.edu, athitsos@uta.edu

1. Implementation Details

In this section, we provide additional details about our experiments for both Breakfast [3] and Hollywood Extended [1] datasets.

In all our experiments, we trained our three proposed networks (Duration, Verb and Object Selectors) together with a dropout value of 0.89 and L2 regularization coefficient of 0.0001 for 40 epochs when using [2] as our pseudo ground-truth, and 90 epochs when using [5] and [4] pseudo ground-truth. Our input features were sampled every three frames over $\alpha = 60$ frames, at the start of each segment in time.

1.1. The Breakfast Dataset Experiments

We set 19 and 14 to be the number of objects and verbs (including background as a separate object/verb) in the Breakfast dataset. ζ, β, and λ were adjusted to 1, 30, and 5 respectively for our selector network using [5] and [4] as the baseline. In experiments where TCFPN results [2] were used as the initial pseudo ground-truth, the aforementioned parameters were slightly changed to 1, 40, and 1.

1.2. The Hollywood Dataset Experiments

There are 17 actions (including the background) in the Hollywood Extended dataset, and most of these actions do not share verbs or objects with each other. Hence, it would be redundant to decompose the main actions into their verb and object attributes. As a result, for this dataset, we removed the object selector component and used the 17 main actions as our verbs. β, and λ were set to 3 and 1, and 20 and 1 for the TCFPN [2] and NNViterbi [5] baselines respectively. In cases where CDFL [4] were used, β was increased to 50.

Around 60% of the frames are background in this dataset. Therefore, it is worth mentioning that a naive classifier, that outputs "background" for every single frame, can achieve results competitive to the state-of-the-art on the *acc* metric. This is why we emphasize that, specifically for the Hollywood Extended dataset, evaluation using *acc-bg* is more informative. Our method outperforms existing models on this metric while producing better or competitive results on *IoU*.

1.3. Competitors' Results

During our observations, we realized that the provided frame-level features are missing for a significant amount of frames in four videos[1] in the Breakfast dataset. While TCFPN [2], NNViterbi [5] and CDFL [4] originally trimmed those videos, we decided to remove them for all experiments including our method as well as all baselines [2, 5, 4]. In Tables 1 and 2 of the main paper, we denote with symbol † the best results that we obtained after running the authors' source code for multiple times. The reason we ran the code multiple times is that each training process is randomly initialized and leads to different final result.

For CDFL [4] in Table 1 and 2, the alignment *acc-bg* on the Hollywood dataset is somewhat different than the one mentioned in the referenced paper. Similarly, for TCFPN [2], in some cases, our reproduced results are not the same as the ones mentioned in [2]. In this case, we reported the results after contacting the authors and having their approval. For a fair comparison in both baselines, we reported the results, that represent the initial pseudo ground-truth in our method.

Without loss of generality, our final accuracy depends on the quality of the initial pseudo ground-truth, so we have provided the initial pseudo ground-truth and pre-trained main action recognizer models (for TCFPN and NNViterbi on the Breakfast dataset) that we used as supplementary material so our results can be reproduced precisely. All the code and pre-trained models that we provide in supplementary material will be publicly available upon publication of this paper.

[1] 1-P34_cam01_P34_friedegg, 2-P51_webcam01_P51_coffee, 3-P52_stereo01_P52_sandwich, 4-P54_cam01_P54_pancake

References

[1] Piotr Bojanowski, Rémi Lajugie, Francis Bach, Ivan Laptev, Jean Ponce, Cordelia Schmid, and Josef Sivic. Weakly supervised action labeling in videos under ordering constraints. In *European Conference on Computer Vision*, pages 628–643. Springer, 2014.

[2] Li Ding and Chenliang Xu. Weakly-supervised action segmentation with iterative soft boundary assignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6508–6516, 2018.

[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.

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

[5] Alexander Richard, Hilde Kuehne, Ahsan Iqbal, and Juergen Gall. Neuralnetwork-viterbi: A framework for weakly supervised video learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7386–7395, 2018.