Supplementaty Material for: "FIFA: Fast Inference Approximation for Action Segmentation"

Yaser Souri[1], Yazan Abu Farha[1], Fabien Despinoy[2], Gianpiero Francesca[2], and Juergen Gall[1]

[1] University of Bonn
[2] Toyota Motor Europe
{souri, abufarha, gall}@iai.uni-bonn.de
{fabien.despinoy, gianpiero.francesca}@toyota-motor.com

1 Detials of Exact Inference

The NNV approach [10] proposes an exact solution to the inference problem using a Viterbi-like dynamic programming method. This dynamic programming approach was later adopted by CDFL [8] and MuCon [11]. First an auxiliary function $Q(t, \ell, n)$ is defined that yields the best probability score for a segmentation up to frame t satisfying the following conditions:

- the length of the last segment is ℓ,
- the last segment was the nth segment with label c_n.

The function Q can be computed recursively. The following two cases are distinguished. The first case defines when no new segment is hypothesized, i.e $\ell > 1$. Then,

$$Q(t, \ell, n) = Q(t-1, \ell-1, n) \cdot p(c_n|x_t), \tag{1}$$

with the current frame probability being multiplied with the value of the auxiliary function at the previous frame. The second case is a new segment being hypothesized at frame t, i.e. $\ell = 1$. Then,

$$\begin{aligned} &Q(t, \ell = 1, n) = \\ &\max_{\hat{\ell}} \left\{ Q(t-1, \hat{\ell}, n-1) \cdot p(c_n|x_t) \cdot p(\hat{\ell}|c_{n-1})) \right\}, \end{aligned} \tag{2}$$

where the optimization being calculated over all possible previous segments with length $\hat{\ell}$ and label c_{n-1}. Here the probability of the previous segment having length $\hat{\ell}$ and label c_{n-1} is being multiplied to the previous value of the auxiliary function.

The most likely alignment is given by

$$\underset{\ell}{max} \left\{ Q(T, \ell, N) \cdot p(\ell|c_N) \right\}. \tag{3}$$

The optimal lengths can be obtained by keeping track of the maximizing arguments $\hat{\ell}$ from (2).

2 Time Complexity Comparison

2.1 Time Complexity of Exact Inference

The time complexity of the above exact inference is quadratic in the length of the video T and linear in the number of segments N. As input videos for action segmentation are usually long, it becomes computationally expensive to calculate. In practice, [10,8,11] limit the maximum size of each segment to a fixed value of $L = 2000$. The final time complexity of exact inference is $O(LNT)$. Furthermore, this optimization process is inherently not parallelizable. This is due to the max operation in (2). Experiments have shown [11,12] that this inference stage is the main computational bottleneck of action segmentation approaches.

2.2 Time Complexity of FIFA

At each optimization step, the time complexity is $O(NT)$, where N is the number of segments and T is the length of the video because we must create the M^* matrix and calculate the element-wise multiplication. Overall, the FIFA time complexity is $O(MNT)$, where M is the number of optimization steps. Compared to the exact inference which has a time complexity of $O(LNT)$, where L is the fixed value of 2000, our time complexity is lower since M is usually 50 steps and N is on average 10.

We also want to mention that the proposed approach is inherently a parallelizable optimization method (i.e. values of the mask, the element-wise multiplication, and the calculation of the gradient for each time step can be calculated in parallel) and is independent of any other time step values. This is in contrast to the dynamic programming approaches where the intermediate optimization values for each time step depend on the value of the previous time steps.

3 Details of the Datasets

The **Breakfast** dataset [7] is the most popular and largest dataset typically used for action segmentation. It contains more than 1.7k videos of different cooking activities. The dataset consists of 48 different fine-grained actions. In our experiments, we follow the 4 train/test splits provided with the dataset and report the average.

The **Hollywood extended** dataset [2] contains 937 videos taken from Hollywood movies. The videos contain 16 different action classes. We follow the train/test split strategy of [4,10,8].

The main performance metrics used for weakly supervised action segmentation and alignment are the same as the previous approaches. The input features are also kept the same depending on the approach we use FIFA with.

Num. Steps	MoF	MoF-BG	IoU	IoD	Time (min)
No inference	45.4	44.7	37.3	51.2	1.0
2 steps	47.9	47.1	39.8	53.0	1.2
5 steps	49.1	48.3	40.0	52.8	1.5
10 steps	50.1	49.4	40.2	52.9	2.0
30 steps	51.2	50.6	41.0	53.2	4.2
50 steps	**51.3**	**50.7**	**41.1**	**53.3**	6.5
60 steps	**51.3**	**50.7**	**41.1**	**53.3**	7.7
Exact Inference	50.7	50.3	40.9	54.0	32.85

Table 1. Impact of the number of optimization steps for FIFA+MuCon for weakly supervised action segmentation on the Breakfast dataset.

4 Implementation Details

We implement our approach using the PyTorch [9] library. For all experiments we set the number of FIFA's gradient-based optimization steps to 50 and we use the Adam [6] optimizer. Mask sharpness and the optimization learning rate is chosen depending on the approach that FIFA is applied on top of. When applying FIFA on top of MuCon [11] we use 0.3 as the learning rate and set the mask sharpness to 1.75. For CDFL [8], we set the mask sharpness to 0.1 and the learning rate to 0.15. Looking at the visualization in Figure 9 it is clear that CDFL provides noisy framewise probability estimates. For this reason a lower mask sharpness is prefered. When applying FIFA on top of fully supervised approaches like MS-TCN [1] we use mask sharpness value of 15 and learning rate of 0.02. Looking at the visualization in Figure 15 we see that fully supervised approaches provide clean smooth framewise probabilities and having a sharp mask is recommented in these settings.

5 Ablation Experiments

5.1 Number of Optimization Steps

In Table 1 we report the results for weakly supervised action segmentation on the Breakfast dataset [7] using the MuCon [11] approach. The proposed approach achieves the best performance after 50 steps with 5.9% improvement on the MoF accuracy compared to not performing any optimization. Moreover, it is more than 5 times faster than the exact inference.

5.2 Optimizer and Its Learning Rate

The choice of the optimizer used to update the length estimates using the calculated gradients is one of the hyper-parameters of our approach. We have experimented with two optimizers SGD and Adam. As shown in Figure 1, the best

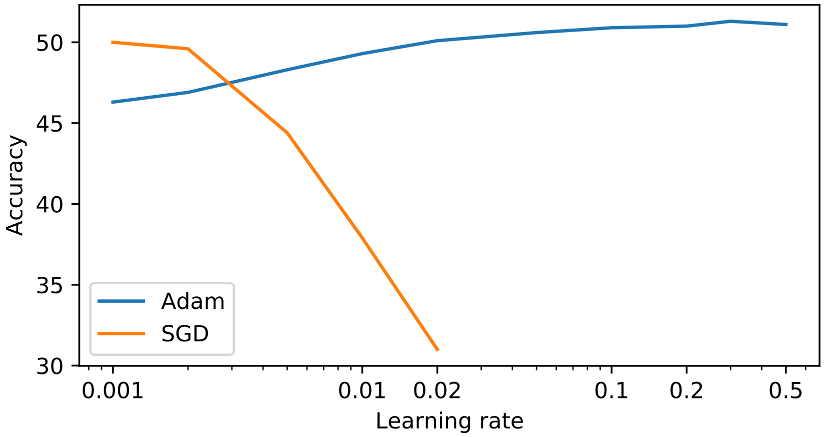

Fig. 1. Effect of the learning rate on the performance of weakly supervised action segmentation using FIFA applied on the MuCon approach. Accuracy is calculated on the Breakfast dataset.

performing value for the learning rate hyper-parameter depends on the optimizer used. For SGD a low value of 0.001 achieves the best performance with higher values causing major drops in performance. On the other hand, Adam optimizer works well with a range of learning rate values as it has an internal mechanism to adjust the learning rate. The best performance for Adam is observed at 0.3.

We further investigate and notice that the reason SGD performs so poorly for large values of the learning rate is that it fluctuates and is not able to optimize the energy effectively. Figure 2 shows the value of the approximate energy during the optimization for Adam and SGD for the same inference. We observe that a large learning rate causes SGD to fluctuate while Adam is stable and achieves a lower energy value at the end of the optimization.

6 Weakly Supervised Action Alignment

Similar to weakly supervised action segmentation, we apply FIFA on top of CDFL and MuCon for weakly supervised action alignment on the Breakfast dataset and report the results in Table 2. Our experiments show that FIFA applied on top of CDFL achieves state-of-the-art or better than state-of-the-art results on MoF and Mof-BG metrics, whereas FIFA applied on top of MuCon achieves state-of-the-art results for IoD and IoU metrics.

7 Qualitative Examples

In this section we show various qualitative results of applying FIFA for action segmentation. In each figure on the right, the approximate total energy value is

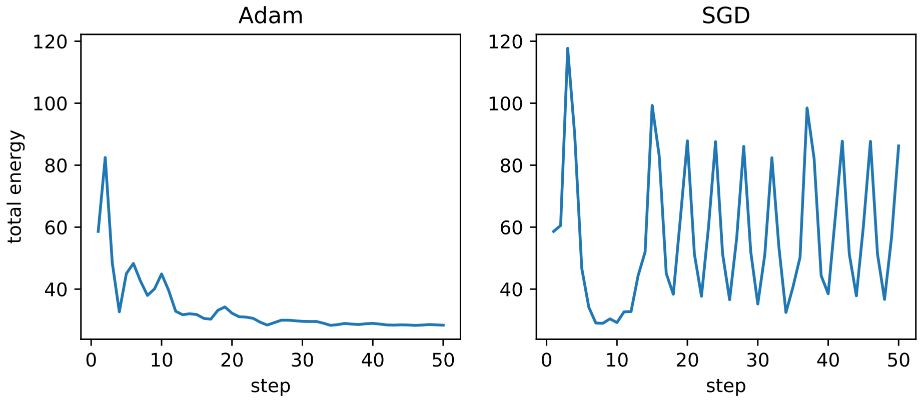

Fig. 2. The value of the approximate energy during FIFA optimization for SGD and Adam optimizer for the same inference.

Method	MoF	MoF-BG	IoU	IoD
ISBA [4]	53.5	51.7	35.3	52.3
D^{3}TW [3]	57.0	-	-	56.3
CDFL [8]	63.0	61.4	45.8	<u>63.9</u>
ADP [5]	<u>64.1</u>	**65.5**	43.0	-
FIFA + CDFL*	**65.3**	<u>64.3</u>	<u>46.3</u>	61.3
FIFA + MuCon*	61.4	61.2	**48.4**	**64.1**

Table 2. Results for weakly supervised action alignment on the Breakfast dataset.

plotted as a function of number of steps. On the left, at the top, the framewise negative log probabilities (P) are visualized. The ground truth segmentation, optimization initialization, the generated masks and the segmentation obtained after approximate inference using FIFA is visualized in rows 2 to 5. The MoF metric is also calculated for a single video and reported for the optimization initialization and the approximate decoding.

An animation of the same figures is also provided in the supplementary material as a single video file.

Figures 3-7 show qualitative examples of applying FIFA on top of MuCon [11] for weakly supervised action segmentation. Figures 9-13 show qualitative examples of applying FIFA on top of CDFL [8] for weakly supervised action segmentation. Figures 15-19 show qualitative examples of applying FIFA on top of MSTCN [1] for fully supervised action segmentation.

7.1 Failure Cases

Figures 8, 14 and, 20 show failure cases for FIFA + MuCon, FIFA + CDFL and, FIFA + MSTCN respectively. We observe that the major failure case is when the optimization is initialized with an incorrect transcript (Figures 8 and 14). Another failure mode is when the predicted negative log probabilities are not correct (Figure 20) which causes the boundaries of actions to be in the wrong location.

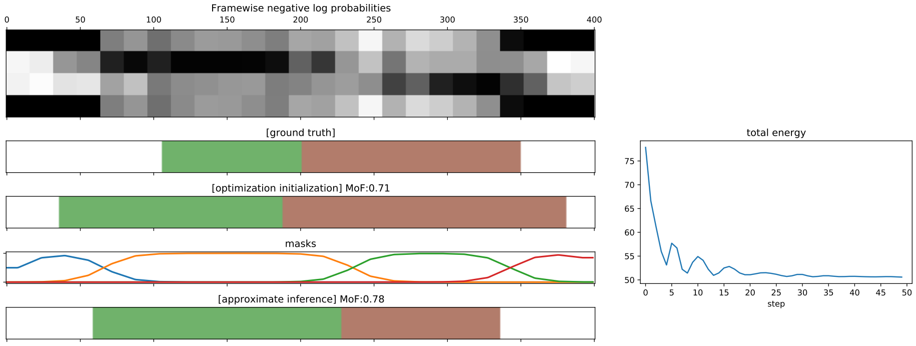

Fig. 3. Qualitativ Result: Weakly supervised action segmentation, FIFA + MuCon

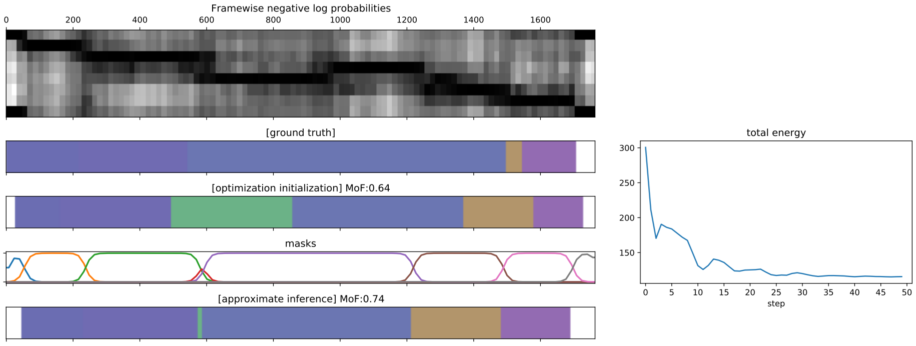

Fig. 4. Qualitativ Result: Weakly supervised action segmentation, FIFA + MuCon

References

1. Abu Farha, Y., Gall, J.: Ms-tcn: Multi-stage temporal convolutional network for action segmentation. CVPR (2019)

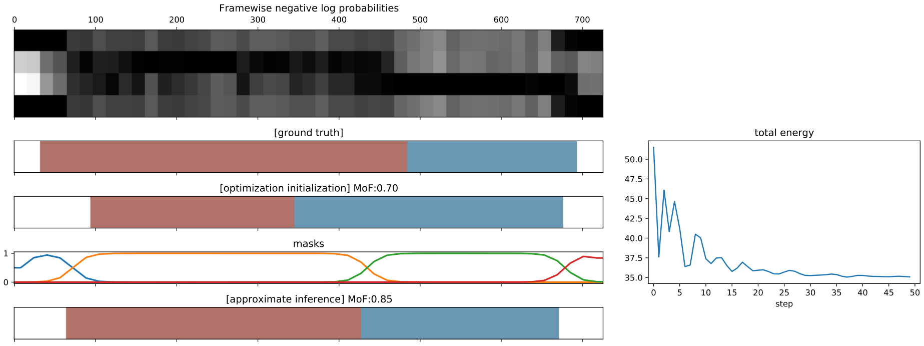

Fig. 5. Qualitativ Result: Weakly supervised action segmentation, FIFA + MuCon

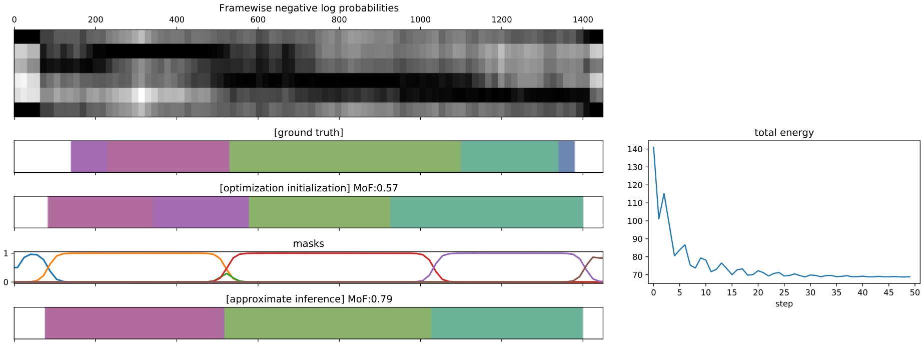

Fig. 6. Qualitativ Result: Weakly supervised action segmentation, FIFA + MuCon

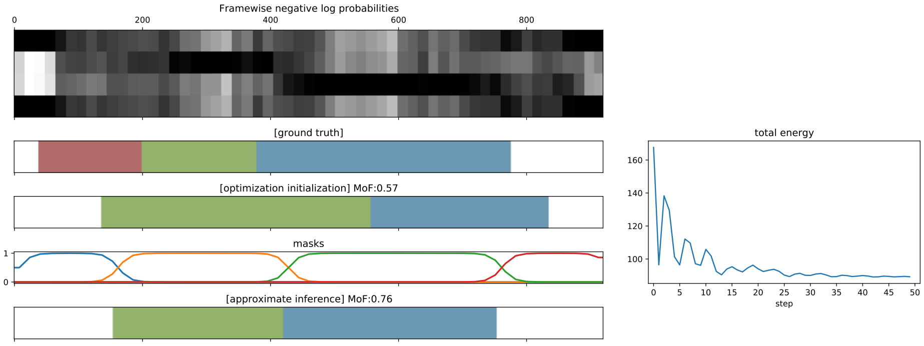

Fig. 7. Qualitativ Result: Weakly supervised action segmentation, FIFA + MuCon

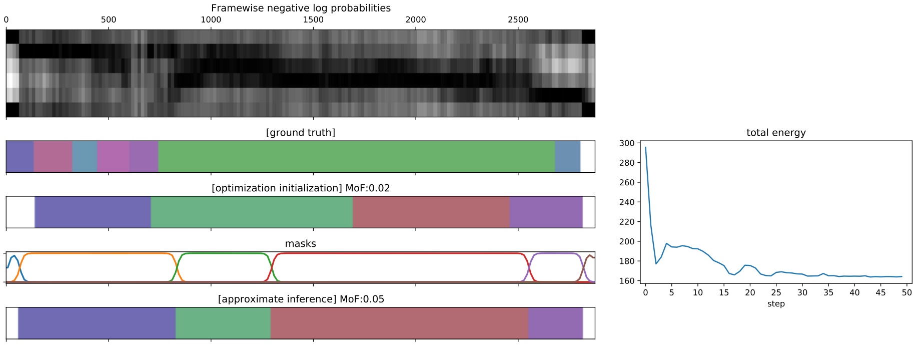

Fig. 8. Qualitativ Result: Weakly supervised action segmentation, FIFA + MuCon, Failure Case

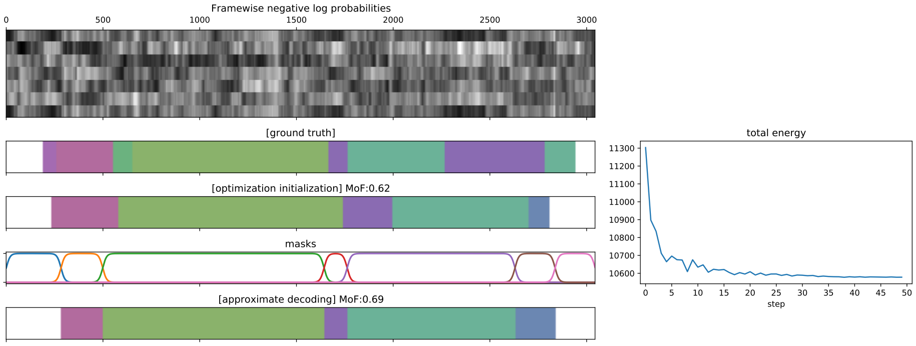

Fig. 9. Qualitativ Result: Weakly supervised action segmentation, FIFA + CDFL

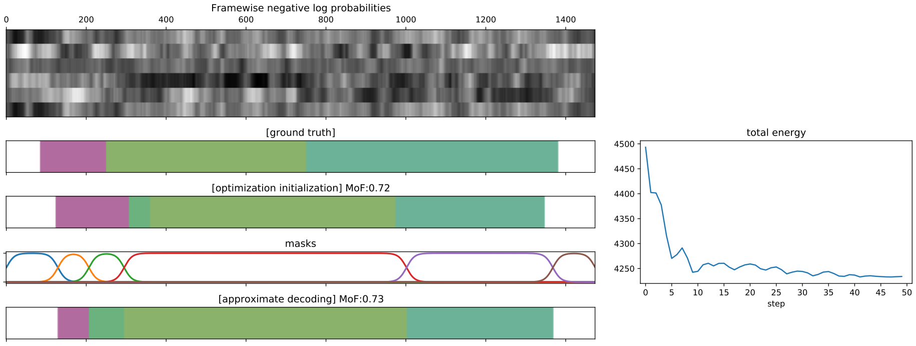

Fig. 10. Qualitativ Result: Weakly supervised action segmentation, FIFA + CDFL

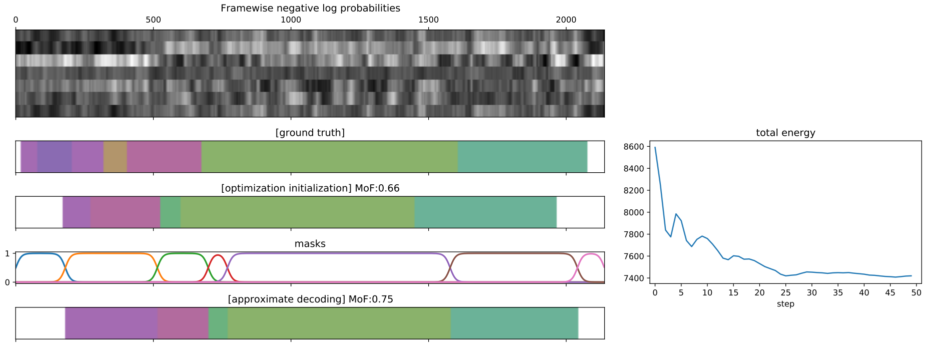

Fig. 11. Qualitativ Result: Weakly supervised action segmentation, FIFA + CDFL

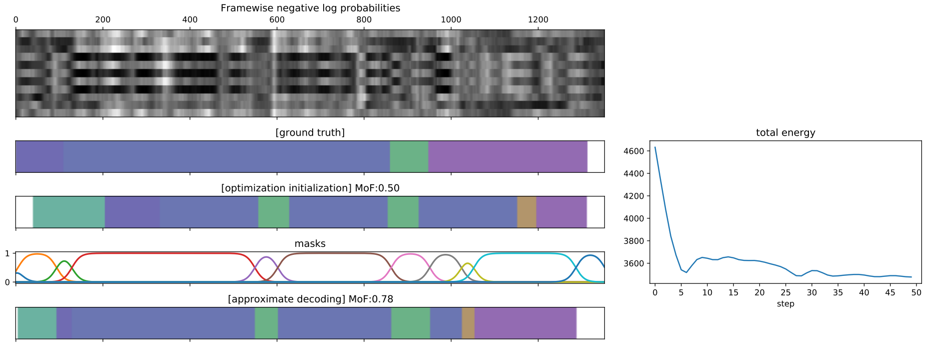

Fig. 12. Qualitativ Result: Weakly supervised action segmentation, FIFA + CDFL

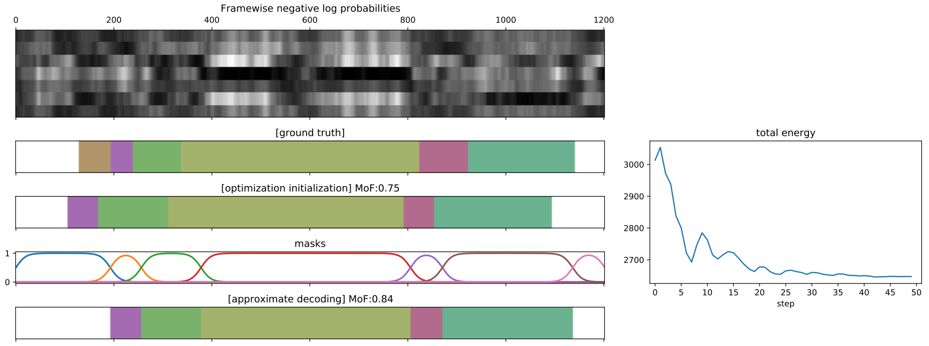

Fig. 13. Qualitativ Result: Weakly supervised action segmentation, FIFA + CDFL

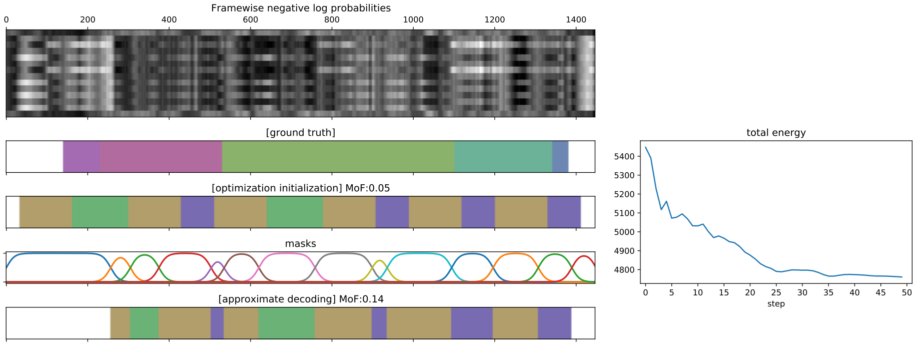

Fig. 14. Qualitativ Result: Weakly supervised action segmentation, FIFA + CDFL, Failure Case

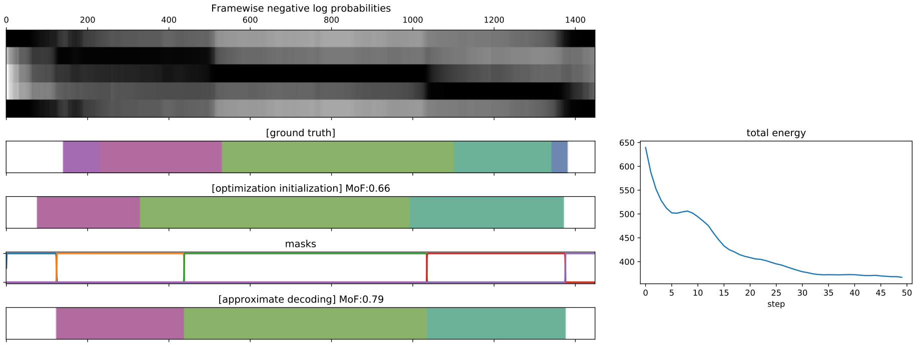

Fig. 15. Qualitativ Result: Fully supervised action segmentation, FIFA + MSTCN

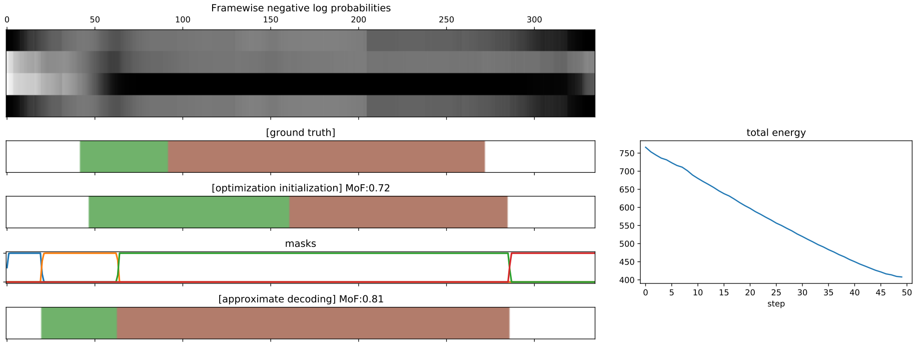

Fig. 16. Qualitativ Result: Fully supervised action segmentation, FIFA + MSTCN

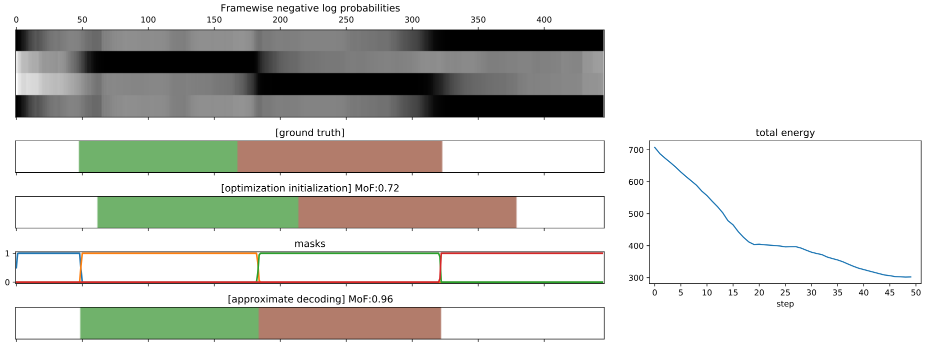

Fig. 17. Qualitativ Result: Fully supervised action segmentation, FIFA + MSTCN

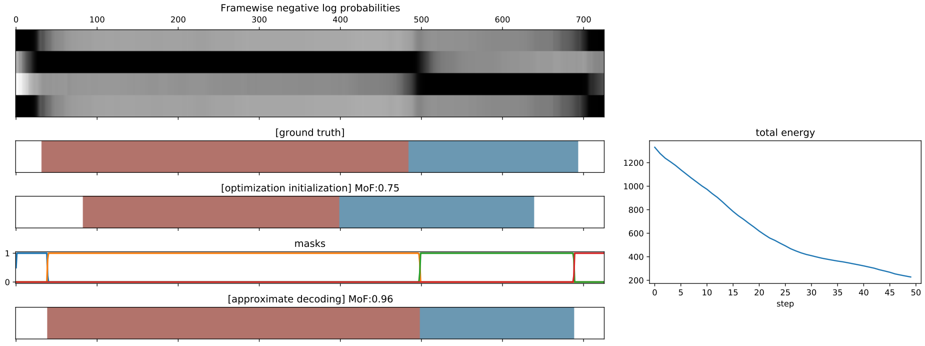

Fig. 18. Qualitativ Result: Fully supervised action segmentation, FIFA + MSTCN

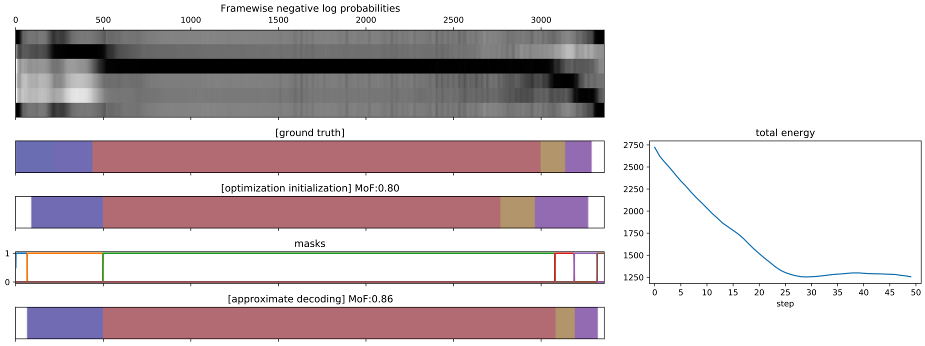

Fig. 19. Qualitativ Result: Fully supervised action segmentation, FIFA + MSTCN

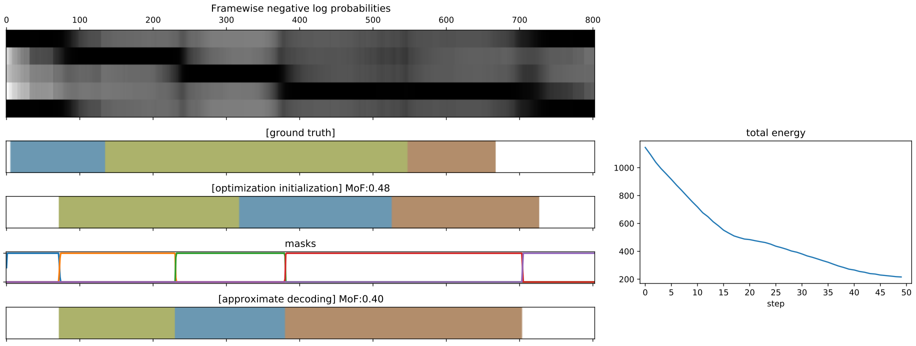

Fig. 20. Qualitativ Result: Fully supervised action segmentation, FIFA + MSTCN, Failure Case

2. Bojanowski, P., Lajugie, R., Bach, F., Laptev, I., Ponce, J., Schmid, C., Sivic, J.: Weakly supervised action labeling in videos under ordering constraints. In: ECCV (2014)
3. Chang, C., Huang, D., Sui, Y., Fei-Fei, L., Niebles, J.C.: D^{3}TW: Discriminative differentiable dynamic time warping for weakly supervised action alignment and segmentation. CVPR (2019)
4. Ding, L., Xu, C.: Weakly-supervised action segmentation with iterative soft boundary assignment. In: CVPR (2018)
5. Ghoddoosian, R., Sayed, S., Athitsos, V.: Action duration prediction for segment-level alignment of weakly-labeled videos. In: WACV (2021)
6. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)
7. Kuehne, H., Arslan, A., Serre, T.: The language of actions: Recovering the syntax and semantics of goal-directed human activities. In: CVPR (2014)
8. Li, J., Lei, P., Todorovic, S.: Weakly supervised energy-based learning for action segmentation. In: ICCV (2019)
9. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: NeurIPS (2019)
10. Richard, A., Kuehne, H., Iqbal, A., Gall, J.: Neuralnetwork-viterbi: A framework for weakly supervised video learning. In: CVPR (2018)
11. Souri, Y., Fayyaz, M., Minciullo, L., Francesca, G., Gall, J.: Fast Weakly Supervised Action Segmentation Using Mutual Consistency. PAMI (2021)
12. Souri, Y., Richard, A., Minciullo, L., Gall, J.: On evaluating weakly supervised action segmentation methods. In: arXiv (2020)