Actor-centered Representations for Action Localization in Streaming Videos

Anonymous ECCV submission

Paper ID 5764

1 Localization Algorithm

Due to space constraints, we could not present a pseudo-code for the localization algorithm. We outline the process in Algorithm 1. The input to the localization process consists of (i) initial regions of interests generated based on spatial features $\mathcal{B}_t^S$ (from Section 3.1 in the main paper), (ii) the spatial-temporal prediction error $\mathcal{L}_{event}$ (from Section 3.2), (iii) number of attention "grids" to consider K, and (iv) the total number of bounding box predictions per frame t.

Algorithm 1: Attention-based Action Localization

Input : $\mathcal{B}_t^S, \mathcal{L}_{event}, K, N$
Output: $\mathcal{B}_t^E$
Initialize: $\mathcal{B}_t^E \leftarrow \emptyset$
$\alpha_t^E = softmax(\mathcal{L}_{event})$
while $|\mathcal{B}_t^E| \leq N$ **do**
  for $e_{i,j} \in \{sorted(\alpha_t^E)\}_{k=0}^{K}$ **do**
    for $b_i \in \mathcal{B}_t^S$ **do**
      if $\mathbb{1}^{obj}(e_{i,j}, b_i)$ **then**
        $\mathcal{B}_t^E \leftarrow \mathcal{B}_t^E \bigcup \{b_i\}$
      end
    end
  end
end

2 Evaluation at various overlap thresholds

We present comparison with state-of-the-art approach at various overlap thresholds on the UCF-Sports dataset in Figure 1. We also compare to PredLearn [2] with two different settings - when the number of clusters is set to the groundtruth K_{gt} and when using the optimal number of clusters K_{opt}, which is typically $3 \times K_{gt}$. It can be seen that we outperform all baselines, even at higher overlap thresholds, including fully supervised models. It is interesting to note that

our approach at K_{gt} outperforms the closely related PredLearn at K_{opt} showing that the use of object-centric representations and the hierarchical prediction helps maintain context when faced with complex motion, both from the camera and the object of interest, to learn robust representations *and* localize the object.

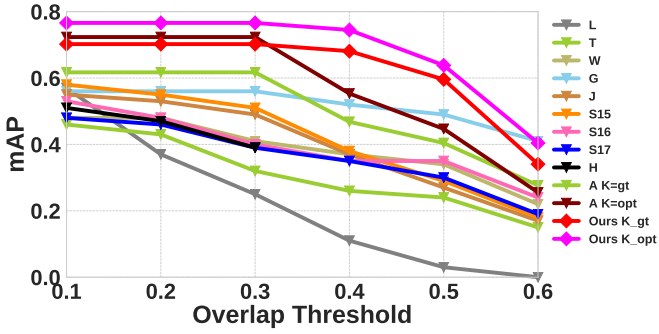
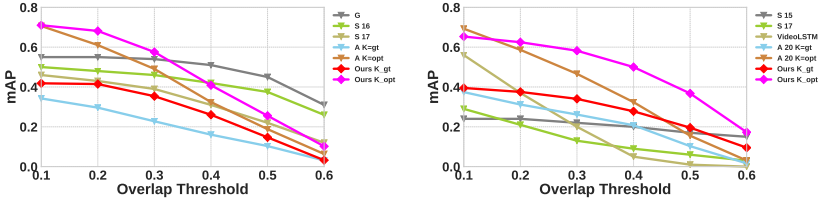


Fig. 1. Comparison with state of the art approaches on the UCF Sports dataset over multiple thresholds. We compare against baselines with varying levels of supervision such as **L**an *et al.* [11], **T**ian *et al.* [21], **W**ang *et al.* [23], **G**kioxari and Malik [5], **J**ain *et al.* [8], **S**oomro *et al.* [17–19], **H**ou *et al.* [7], VideoLSTM [13], and **A**akur *et al.* [2].

We present a similar comparison for the JHMDB and THUMOS'13 datasets in Figure 2. Again, we can see that when using optimal number of clusters we significantly outperform other baselines while using the groundtruth number of classes, we still perform competitively with fully supervised while outperforming other unsupervised and weakly supervised baselines even at higher overlap thresholds.



(a) (b)

Fig. 2. Comparison with state of the art approaches on the (a) JHMDB and (b) THUMOS'13 datasets. We report the mAP over multiple thresholds and compare with several strong baselines.

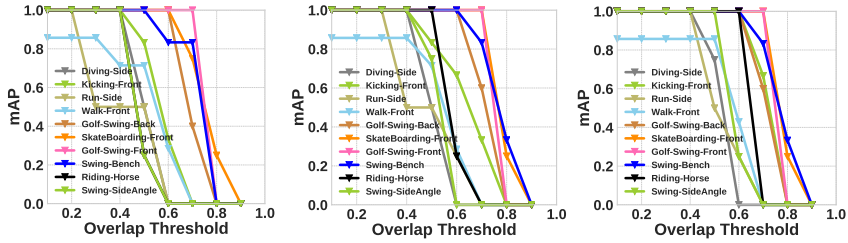


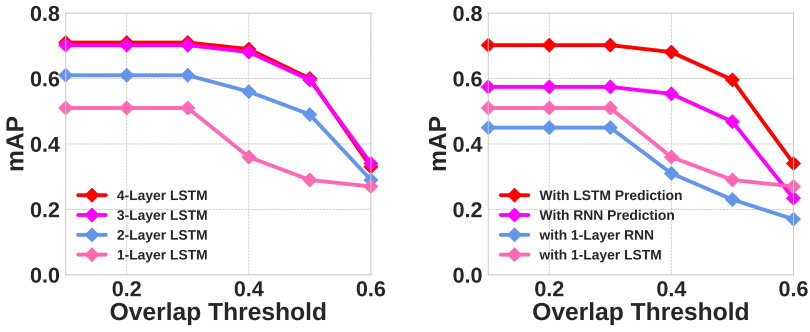
Fig. 3. Class-wise mAP visualized at various overlap thresholds for (a) K_{attn}=1, (b) K_{attn}=5 and (c) K_{attn}=10 on the UCF-Sports dataset.

3 Effect of K_{attn} on different actions

We also present a more qualitative analysis of the use of multiple attention grids K_{attn} (from Section 3.5) on different action classes in the UCF Sports dataset in Figure 3. It can be seen that the use of multiple attention grids has a considerable effect on classes with more background motion such as Walk, Kick and Golf-Swing which have other actors performing similar or distracting actions with complex motion. The use of multiple attention grids allow the model to maintain context in prediction and hence not get distracted from the object of interest, which was the case in PredLearn [2].

4 Effect of LSTM-based Prediction

We analyze the effect of using different configurations in the event-level prediction stack (Section 3.2) in this section. We perform two different ablation studies to find (i) the effect of changing the number of LSTM layers ℓ, and (ii) effect of changing the prediction function from LSTM to RNN. We summarize the results in Figure 4. As can be seen from Figure 4(a), as we reduce the number of LSTM layers from 3, the performance drops drastically, but increasing it to 4 shows very negligible performance improvement. We find that further increasing the number of layers (beyond 4) does not improve the performance and even degrades a little. This could arguably be attributed to the fact that we only use 1 epoch of training and adding more layers causes the model to underfit, leading to performance degradation. We also change the prediction model from an LSTM to an RNN to evaluate the effect of using an explicit "event model" as considered in PredLearn [2] and other continual predictive learning models [1] and summarize the results in Figure 4(b). It can be seen that using the LSTM state as the event model improves the performance of the approach, more significantly in the lower detection thresholds.



(a) (b)

Fig. 4. Ablation studies on the effect of varying the LSTM-based parameters. (a) shows the effect of changing the number of LSTM layers in the prediction stack, and (b) shows the effect of using RNNs instead of LSTMs in the prediction stack.

5 Effect of Training Sequence

The proposed model is continually in a streaming manner. Hence, each training video is presented sequentially to the model and the parameters are updated *every frame*. Note that this is significantly different from most other approaches which are trained in batch mode with each batch containing examples a balanced distribution across classes. In this section, we evaluate the effect of training order on the approach and evaluate its incremental learning capabilities. Specifically, we present the videos *per class* in sequence to the model while training, instead of shuffling it randomly as is the usual practice. We summarize the results in Figure 5. It can be seen that the order of training videos does not have a detrimental effect on the performance of the model. This could arguable be attributed to both the adaptive learning rate scheduling as well as the inherent nature of predictive learning that aims to capture the intra-event correlations and inter-event variations. In fact, at higher thresholds ($\sigma \geq 0.5$), the difference is very negligible or even exceeds that of the random sequence training by a small margin (1.25%).

6 Quality of localizations.

We also independently assess the quality of the localization returned by the approach by computing the *recall* of the bounding boxes returned. We summarize the results in Table 1. Again, it can be seen that the approaches with full supervision (top half) have a higher recall at a more stringent overlap threshold of σ=0.5 while the recall and mAP drop with decreasing levels of supervision. However, it is interesting to note that our approach has a higher recall than both weakly supervised and unsupervised baselines at higher overlap thresholds

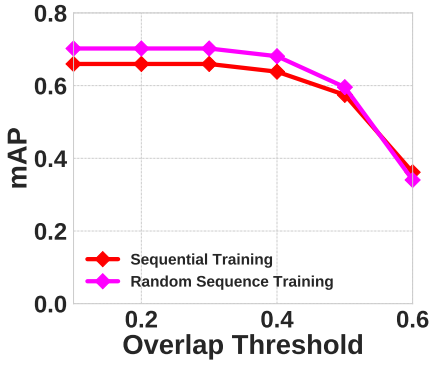


Fig. 5. Ablation study conducted to evaluate how the order of training video sequences affect the continual learning aspect of the proposed approach.

Approach	Average Recall			mAP
	0.1	0.3	0.5	@0.2
Action Tubelets [9]	-	-	0.33	**0.48**
Learning to Track [25]	-	-	**0.61**	0.47
ALSTM [16]	0.46	0.05	-	0.06
VideoLSTM [13]	0.71	0.32	-	0.37
Actor Supervision [3]	**0.89**	-	**0.44**	**0.46**
PredLearn [2]	0.84	0.58	0.33	0.31 (0.59*)
Ours	**0.86**	**0.74**	**0.56**	**0.38** (0.63*)

Table 1. *Quality of localization* on THUMOS'13. We report average recall at various overlap thresholds and the mAP at 0.2 overlap. * refers to evaluation with optimal clusters $k = k_{opt}$.

while also significantly improving (7% absolute mAP) upon prior unsupervised localization approaches.

7 Generalization for Action Recognition

We also evaluate the proposed approach's representation learning ability for the *action recognition* task. We use the first split of UCF-101 [20] and HMDB-51 [10] datasets as the evaluation data following prior work in Motion Statistics [22]. We summarize the results in Table 2 and compare against *early* self-supervised approaches. We evaluate the performance under two conditions, (i) when the number of clusters is set to the ground truth number of classes (k_{gt}) and (ii) allow for over-segmentation by setting k=2k_{gt}. Note that we do not *finetune* on any data and use the model trained on THUMOS'13 to obtain video-level features for clustering. It can be seen that we can learn robust representations that allow

Approach	HMDB51	UCF-101
Invariant Mapping [6]	13.4	38.4
Temporal Coherence [15]	15.9	45.4
Object patch [24]	15.6	42.7
Shuffle & Learn. [14]	19.8	50.9
OPN [12]	22.1	56.3
Geometry [4]	23.3	55.1
Motion Statistics [22]	**32.6**	**58.8**
PredLearn [2]	19.9 (*26.2*)	21.3 (*33.7*)
Ours	**23.6** (*40.4*)	**29.1** (*50.73*)

Table 2. Evaluation of the actor-centered features for recognition. Methods in the bottom *are not finetuned with labeled data.*

the model to cluster the videos in more complex datasets with significantly more classes to a reasonable level while obtaining competitive performance with models *finetuned* on the domains when allowed to over-segment. We see that the clusters' homogeneity score was 79% for UCF-101 and 63% when allowed to over-segment i.e. setting k=2k_{gt}. This indicates that although the number of clusters is higher than the ground-truth, the videos in the same cluster were mostly from the same label.

References

1. Aakur, S.N., Sarkar, S.: A perceptual prediction framework for self supervised event segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
2. Aakur, S.N., Sarkar, S.: Action localization through continual predictive learning. arXiv preprint arXiv:2003.12185 (2020)
3. Escorcia, V., Dao, C.D., Jain, M., Ghanem, B., Snoek, C.: Guess where? actor-supervision for spatiotemporal action localization. Computer Vision and Image Understanding **192**, 102886 (2020)
4. Gan, C., Gong, B., Liu, K., Su, H., Guibas, L.J.: Geometry guided convolutional neural networks for self-supervised video representation learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5589–5597 (2018)
5. Gkioxari, G., Malik, J.: Finding action tubes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 759–768 (2015)
6. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). vol. 2, pp. 1735–1742. IEEE (2006)
7. Hou, R., Chen, C., Shah, M.: Tube convolutional neural network (t-cnn) for action detection in videos. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 5822–5831 (2017)
8. Jain, M., Van Gemert, J., Jégou, H., Bouthemy, P., Snoek, C.G.: Action localization with tubelets from motion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 740–747 (2014)

9. Jain, M., Van Gemert, J., Jégou, H., Bouthemy, P., Snoek, C.G.: Tubelets: Unsupervised action proposals from spatiotemporal super-voxels. International Journal of Computer Vision **124**(3), 287–311 (2017)
10. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: Hmdb: a large video database for human motion recognition. In: 2011 International Conference on Computer Vision. pp. 2556–2563. IEEE (2011)
11. Lan, T., Wang, Y., Mori, G.: Discriminative figure-centric models for joint action localization and recognition. In: 2011 International Conference on Computer Vision. pp. 2003–2010. IEEE (2011)
12. Lee, H.Y., Huang, J.B., Singh, M., Yang, M.H.: Unsupervised representation learning by sorting sequences. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 667–676 (2017)
13. Li, Z., Gavrilyuk, K., Gavves, E., Jain, M., Snoek, C.G.: Videolstm convolves, attends and flows for action recognition. Computer Vision and Image Understanding **166**, 41–50 (2018)
14. Misra, I., Zitnick, C.L., Hebert, M.: Shuffle and learn: unsupervised learning using temporal order verification. In: European Conference on Computer Vision. pp. 527–544. Springer (2016)
15. Mobahi, H., Collobert, R., Weston, J.: Deep learning from temporal coherence in video. In: Proceedings of the 26th Annual International Conference on Machine Learning. pp. 737–744 (2009)
16. Sharma, S., Kiros, R., Salakhutdinov, R.: Action recognition using visual attention. In: Neural Information Processing Systems: Time Series Workshop (2015)
17. Soomro, K., Idrees, H., Shah, M.: Action localization in videos through context walk. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3280–3288 (2015)
18. Soomro, K., Idrees, H., Shah, M.: Predicting the where and what of actors and actions through online action localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2648–2657 (2016)
19. Soomro, K., Shah, M.: Unsupervised action discovery and localization in videos. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 696–705 (2017)
20. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
21. Tian, Y., Sukthankar, R., Shah, M.: Spatiotemporal deformable part models for action detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2642–2649 (2013)
22. Wang, J., Jiao, J., Bao, L., He, S., Liu, Y., Liu, W.: Self-supervised spatio-temporal representation learning for videos by predicting motion and appearance statistics. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4006–4015 (2019)
23. Wang, L., Qiao, Y., Tang, X.: Video action detection with relational dynamic-poselets. In: European Conference on Computer Vision. pp. 565–580. Springer (2014)
24. Wang, X., Gupta, A.: Unsupervised learning of visual representations using videos. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2794–2802 (2015)
25. Weinzaepfel, P., Harchaoui, Z., Schmid, C.: Learning to track for spatio-temporal action localization. In: Proceedings of the IEEE international conference on computer vision. pp. 3164–3172 (2015)