

Embarrassingly Simple Model for Early Action Proposal

M. Baptista-Ríos, R. J. López-Sastre,
F. J. Acevedo-Rodríguez and S. Maldonado-Bascón.

GRAM, Department of Signal Theory and Communications, University of Alcalá
Published in the Anticipating Human Behavior Workshop, ECCV 2018

Abstract. Early action proposal consists in generating high quality candidate temporal segments that are likely to contain an action in a video stream, as soon as they happen. Many sophisticated approaches have been proposed for the action proposal problem but from the off-line perspective. On the contrary, we focus on the *on-line* version of the problem, proposing a simple classifier-based model, using standard 3D CNNs, that performs significantly better than the state of the art.

Keywords: early action proposal, logistic regression, deep learning

Introduction. In this work, we introduce the novel problem of Early Action Proposal (EAP). Unlike traditional off-line activity proposal approaches [1,3,6,4,2], we move towards the on-line version of the problem, where the goal is to generate high quality action candidate temporal segments in a video stream, but as soon as they happen. This novel *early* setting can be useful in many practical applications, where the video arrives in an on-line fashion, such as for robotics, or video surveillance cameras. Moreover, we show that the sophisticated off-line solutions [1,3,6,4,2], which define the current state-of-the-art, offer a poor performance for the EAP problem, mainly because they assume a more simplified setup, where the whole video is always available to produce the proposals.

Model description. As it is shown in Figure 1, for EAP, the action proposal must be generated on-line. This requires identifying whether the action is taking place or not, directly from the video stream, hence working with partial observations of the actions, and ideally with a minimal latency, which is in contrast to the complex sampling mechanisms of most of the off-line models. In addition, an EAP solution must correctly discriminate the action from the background frames, the latter being more frequent.

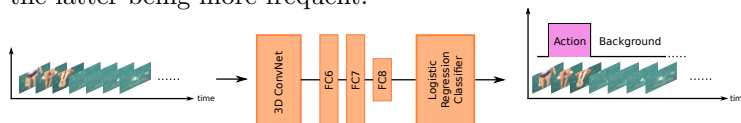


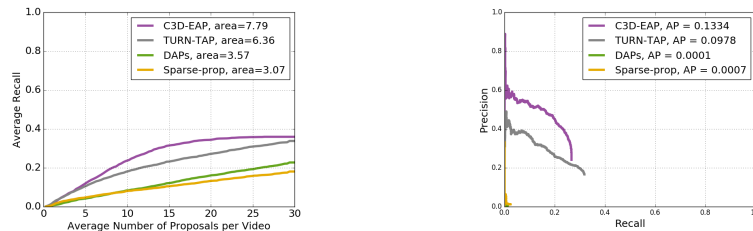
Fig. 1: Proposed C3D-EAP model. We use a logistic regression classifier, with C3D features, trained to discriminate action from background.

We propose a simple solution, which is based on the 3D CNNs (C3D) [7]. We name our model as C3D-EAP. Technically, it consists of learning a C3D network to discriminate between action and background. Then, for each set of frames of a given test video, our network indicates whether they are action or background.

With this output (see Figure 1), we can build the action proposals in an on-line fashion. In order to assign a score for each of our early proposals, we use the mean of the scores, provided by the logistic regression classifier, for the set of evaluated frames of the proposal.

Experiments. For the experimental validation of the EAP problem, we use the untrimmed videos from the THUMOS’14 dataset [5]. We report results on the 213 test videos, using the validation set for learning our approach. We compare our work (C3D-EAP) with those state-of-the-art approaches whose authors provide results, *i.e.* Sparse-Prop [1], DAPs [3], and the recent TURN-TAP [4].

As for the evaluation, we use the standard metric Average Recall at different Average Number of Proposals per Video (AR-AN), used by all the off-line state-of-the-art models [1,3,6,4,2]. However, for the novel EAP problem it is important to also evaluate whether the proposals are likely to include the action of interest, ideally achieving high recall with few proposals. Therefore, as it is shown in Figure 2a, we limit the number of proposals to 30 per video¹, during the evaluation. Furthermore, to measure the quality of the proposals for the EAP problem, we also advocate that it is important to recover the well-known Precision-Recall (PR) curve as an evaluation metric (see Figure 2b).



(a) AR-AN with 0.5 tIoU threshold.

(b) P-R with 0.5 tIoU threshold.

Fig. 2: Comparison with the state-of-the-art on THUMOS’14 [5] dataset.

Figure 2 shows that our embarrassingly simple model clearly outperforms the state-of-the-art approaches for both metrics. While those approaches generate many proposals in an off-line fashion, our solution is able to work on-line, generating fewer proposals, that are also more accurate. This can be seen in the qualitative examples presented in Figure 3.

References

1. Caba-Heilbron, F., Niebles, J.C., Ghanem, B.: Fast temporal activity proposals for efficient detection of human actions in untrimmed videos. In: CVPR (June 2016)
2. Chao, Y.W., Vijayanarasimhan, S., Seybold, B., Ross, D.A., Deng, J., Sukthankar, R.: Rethinking the faster R-CNN architecture for temporal action localization. In: CVPR (2018)
3. Escorcia, V., Caba Heilbron, F., Niebles, J.C., Ghanem, B.: DAPs: Deep action proposals for action understanding. In: ECCV. pp. 768–784 (2016)

¹ In the THUMOS’14 test set, each video contains 15 action instances on average.

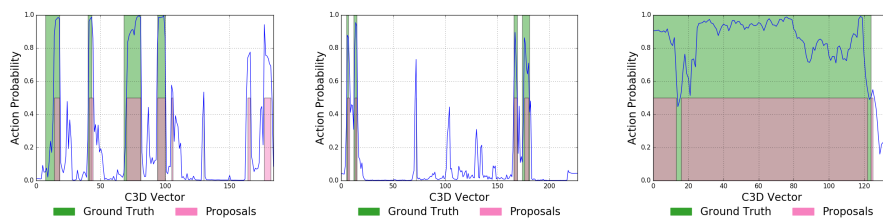


Fig. 3: Qualitative results for our C3D-EAP approach.

4. Gao, J., Yang, Z., Chen, K., Sun, C., Nevatia, R.: TURN TAP: Temporal unit regression network for temporal action proposals. In: ICCV (Oct 2017)
5. Jiang, Y.G., Liu, J., Roshan Zamir, A., Toderici, G., Laptev, I., Shah, M., Suktanar, R.: THUMOS challenge: Action recognition with a large number of classes. <http://crcv.ucf.edu/THUMOS14/> (2014)
6. Shou, Z., Wang, D., Chang, S.F.: Temporal action localization in untrimmed videos via multi-stage cnns. In: CVPR (2016)
7. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: ICCV. pp. 4489–4497 (Dec 2015)