C3D and Localization Model for Locating and Recognizing the Actions from Untrimmed Videos (Student Abstract)

Himanshu Singh, Tirupati Pallewad, Badri N Subudhi, and Vinit Jakhetiya

Indian Institute of Technology Jammu, Jammu & Kashmir, India
2020ree2057@iitjammu.ac.in, 2019pee0037@iitjammu.ac.in, subudhi.badri@iitjammu.ac.in, vinit.jakhetiya@iitjammu.ac.in

Abstract

In this article, we proposed a technique for action localization and recognition from long untrimmed videos. It consists of C3D CNN model followed by the action mining using the localization model, where the KNN classifier is used. We segment the video into expressible sub-action known as action-bytes. The pseudo labels have been used to train the localization model, which makes the trimmed videos untrimmed for action-bytes. We present experimental results on the recent benchmark trimmed video dataset "Thumos14".

Introduction

Action localization and recognition are two important tasks in Computer Vision. Actions in a video are continuous over frames. In a trimmed video, the actions are temporally aligned and are easier to analyze as compared to the untrimmed videos, where along with the actions several background or irrelevant details are also present, which makes it difficult to analyze.

Several pioneer works have been reported by the researchers in the state-of-the-art (SOTA) techniques for action localization and recognition. In conventional practice, two approaches namely: fully supervision-based and weakly supervised classification-based algorithms are quite popular. In the prior method, the complete temporal annotations are trained with a supervised classifier to predict the class labels of different actions in a video. Whereas, in the latter one, the action labels are used for training and prediction of the actions in a video (Jain, Ghodrati, and Snoek 2020). However, short-duration training videos for action recognition are always not effective in detecting accurate action boundaries.

In this work, we devise a novel technique for localization and recognition of actions in long-trimmed videos. In the proposed scheme we adhered to a C3D CNN model followed by the action localization model. C3D CNN architecture provides a generic video feature for action localization in videos that contains different kind of activities. It also provides compact representation as well as better separability across the action classes. The proposed scheme uses the multiple instance learning loss and co-activity similarity loss to classify and localize the actions in an untrimmed video. We evaluate the proposed scheme on untrimmed video database "Thumos14". The evaluation of the proposed scheme is verified using fifteen SOTA techniques and found to be performing very well.

Proposed Methodology

It may be observed that many of the SOTA techniques employ multi-layer CNN architectures for action localization and recognition. However existing CNN approaches are unable to characterize the motion information in a video and incurred complexity. Also fails in the real-time scenario with long untrimmed videos. To resolve the same, we seek the advantages of 3-dimensional convolutional neural network (C3D) architecture to represent both the image-level and motion information in a video. In this article, we propose a unique action localization and recognition network where a pre-trained C3D and a KNN classifier with the localization model are used to classify the actions in a video. The reason behind choosing C3D is, it's efficient capability of capturing the spatio-temporal features from complex activities. We design our model to locate the actions in long untrimmed videos, given the trimmed videos action labels. The architecture of the proposed algorithm is shown in Figure 1. In the proposed scheme, we have initially segmented the video into smaller segments as action-bytes. The proposed architecture is trained using the cluster action-bytes as pseudo-labels. Further, the action-bytes are mined by iterating between generating the pseudo-labels from action-bytes and training the localization model.

Action-bytes The high-level deep features change slowly over time and sudden modifications in the feature space are reflected in the pixel space. We have considered this characteristics to decompose videos into action-bytes. Action-bytes are similar in successive instances from a lengthy video that carries some meaningful information.

We extract d-dimensional features ($D = \{d_t\}_{t=1}^{T_l}$) for every time instant t using a C3D model, where T_l is the temporal length. For simplification and comprehensiveness of the trained model, we transformed the feature space to the latent space. This mapping is applied using a latent projection module, which produces an output vector given by, $I_s \in R^{m \times T_l}$. The latent concepts help in defining the boundaries of action-bytes. Whenever the action changes, there is a significant change in the defined latent concepts. Successive frames are fed to the system to check affinity be-

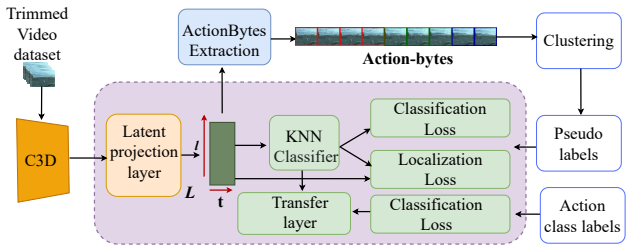

Figure 1: Localization model for extraction of Action-bytes

tween evaluated latent concepts, and the video is divided into sub-classes called as action-bytes. The Action-byte boundary can be given as:

$$Y = \{\sum_{j=1}^{l} |I_s[j, t+1] - I_s[j, t]| > \tau\}, \quad (1)$$

where τ is a predefined constant used for thresholding.

Mining action-bytes We decomposed the video into multiple sub-action videos and then obtained pseudo-labels that we have used to train the classification model. The pipeline for mining action-bytes has two steps: generation of the pseudo-labels from action-bytes, and training of the action localization model using pseudo-labels. It iterates between these two steps. For generating pseudo-labels, we follow Caron et al. (Caron et al. 2018) technique. After extracting N action-bytes from training videos, we take an average of latent features within the duration of each action-byte and represent it in terms of the average value. We adhered to the Partitioning Around Medoids clustering algorithm to cluster all the extracted action-bytes into K clusters.

The main task of our proposed method is to classify the action-bytes and localize them into pseudo-labels. The latent projection module used in the model is simply a fully connected layer followed by ReLU. The output of the above module is passed to a KNN classifier, where we generate the activation scores for pseudo-labels. It provides the predictions for the pseudo-labels. We use a transfer layer on top of the KNN classifier to convert this into projections for the training classes. We have used co-activity similarity loss for localization and k-max multiple-instance learning loss for the classification (Paul, Roy, and Roy-Chowdhury 2018). As we know that multiple instance learning loss and co-activity similarity loss are complementary to each other so we have jointly optimized them to learn our model weights. These weights are used to localize and classify the actions in a given untrimmed video at the time of testing.

Experimental Results

We have tested the proposed technique on the Thumos14 database. We have used two metrics: IoU (Intersection over Union) and mean average precision (mAP), for evaluating our model performance quantitatively. Table 1 shows the performance comparison of the proposed action localization technique using IoU measure against eleven considered SOTA techniques. It may be observed that the proposed technique provides better results than ten SOTA techniques

IoU →	0.1	0.2	0.3	0.4	0.5
Saliency-Pool	04.6	03.4	02.1	01.4	00.9
FV-DTF	36.6	33.6	27.0	20.8	14.4
SLM-mgram	39.7	35.7	30.0	23.2	15.2
S-CNN	47.7	43.5	36.3	28.7	19.0
PSDF	51.4	42.6	33.6	26.1	18.8
R-C3D	54.5	51.5	44.8	35.6	28.9
SSN	**60.3**	**56.2**	**50.6**	**40.8**	**29.1**
HAS	36.4	27.8	19.5	12.7	6.8
UntrimmedNets	44.4	37.7	28.2	21.1	13.7
STPN (UNTF)	45.3	38.8	31.1	23.5	16.2
STPN (I3DF)	52.0	44.7	35.5	25.8	16.9
Proposed Method	**55.55**	**50.95**	**39.81**	**30.29**	**22.54**

Table 1: Evaluation of Proposed Action Localization model with SOTA Techniques on Thumos14 database (SupplementaryMaterials 2021)

Methods	mAP
EMV + RGB	61.5
Objects + Motion	71.6
Temp. Seg. Net. (TSN)	68.5
Two Stream	66.1
Proposed Method	**96.10**

Table 2: Evaluation of the Action classification With SOTA Techniques on Thumos14 database (SupplementaryMaterials 2021)

and comparative results with the Structured Segment Network (SNN) technique. In Table 2 the action recognition accuracy of the proposed algorithm is compared with four SOTA techniques, where the proposed technique is performing the best. The references for all the considered techniques used for comparison are available at (SupplementaryMaterials 2021).

Conclusions

We propose a C3D and localization model for action localization and recognition. The efficiency of the proposed scheme is verified by comparing it against SOTA techniques. Furthermore, the proposed system's ability to handle unknown classes is one of the significant aspects. To improve its accuracy, further tuning of the parameters is in progress.

References

Caron, M.; Bojanowski, P.; Joulin, A.; and Douze, M. 2018. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 132–149.

Jain, M.; Ghodrati, A.; and Snoek, C. G. 2020. ActionBytes: Learning from Trimmed Videos to Localize Actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1171–1180.

Paul, S.; Roy, S.; and Roy-Chowdhury, A. K. 2018. W-talc: Weakly-supervised temporal activity localization and classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 563–579.

SupplementaryMaterials. 2021. :[Online] Available. https://github.com/cdshmnsh/Supplementary-material-for-AAAI. Accessed: 16/10/2021.