Demo: Temporal Action Localization in Untrimmed Videos

Yizhou Wang
Columbia University
New York, NY 10027
wang.yizhou@columbia.edu

Zheng Shou
Columbia University
New York, NY 10027
zheng.shou@columbia.edu

Shih-Fu Chang
Columbia University
New York, NY 10027
shih.fu.chang@columbia.edu

ABSTRACT

Temporal action localization (TAL) is an active research direction that has been widely concerned. Recently, several great frameworks are proposed to solve TAL problems, such as Segment-CNN and CDC Networks. However, demonstration of TAL results is challenging since video segments are hard to represent. To solve the problem, the snippet-level and frame-level demonstration methods are proposed and developed as a web interface, where users can either upload video or select video from THUMOS to processing TAL algorithms. The demonstration methods we proposed can give users TAL results clearly and efficiently.

1 INTRODUCTION

Temporal action localization (TAL) in videos has attracted many researchers in the computer vision community because of its wide range of applications and promising commercial values [1, 3, 7]. The key problem of TAL is not only to recognize what actions are contained in a video, but also to tell the time period of the actions.

Recently, several models to solve TAL problems are proposed. One of them called Segment-CNN (S-CNN) [5] is an end-to-end deep learning framework based on 3D ConvNets [6]. After Segment-CNN, Convolutional-De-Convolutional Networks (CDC) [4] is proposed, which places CDC filters on top of 3D ConvNets and predicts action at the frame-level. Both of these two models obtained good performance on THUMOS 2014 [2].

Since action localization problem concentrate on both the overall and partial information and their relationship, it can be analogized with object detection in images. But demonstration of TAL results is not as easy as object detection in images, because video segments are hard to represent.

Therefore, we proposed a snippet-level demo for SCNN framework, and a frame-level demo for CDC framework, then implemented them on our project website. Firstly, consider a test video is passed through SCNN framework, the results would involve segments with action names, start/end times and confidence scores. To demonstrate the results, we generated a sequence of snippet uniformly and displayed the results for each snippet. In the snippet sequence, people can also easily find out the position of the detected segments in the overall video. Secondly, when a video is passed through CDC framework, the results would be the per-frame action names and scores. Besides the snippet sequence, the results of each frame are also displayed while the video is playing. A short video of our demo can be found at https://youtu.be/wBjiOOzVDDM.

2 DEMO INSTRUCTIONS

The snippet-level demo and the frame-level demo use the same user interface. The interface (Figure 1) contains four parts:

- Frame player: Showing the frames in the test video.

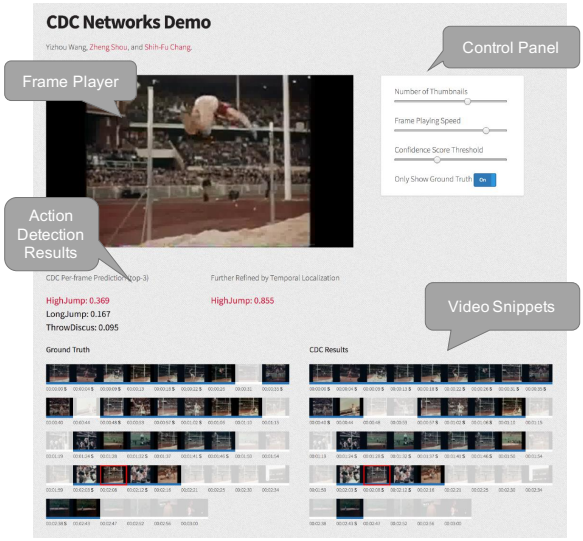


Figure 1: The user interface of the snippet-level demo and the frame-level demo.

- Action detection results: Showing both per-frame and refined predictions and corresponding confidence scores in descending order.
- Video snippets: Each thumbnail represents for a snippet of the video. The color bars beneath represent for the action included/detected in this snippet.
- Control panel: Several parameters can be adjusted here.

2.1 Parameter Settings

In the control panel, there are overall 4 parameters can be adjusted to make the demo more user-friendly:

- `Number of Thumbnails`: With a big number of thumbnails, we can look into the details of a certain segment.
- `Frame Playing Speed`: This is one of the properties of the frame player. It changes the playing speed of the video.
- `Confidence Score Threshold`: It is used to filter out the segments with low confidence scores.
- `Only Show Ground Truth`: It is a switch to filter out the segments without the action(s) in the ground truth.

2.2 Snippet-level Demonstration

The snippet-level method is designed for the segment-level TAL frameworks, such as SCNN framework. The key of snippet-level method is "snippet". The snippet sequence forms a time line of the test video, and masks with different colors represents for different

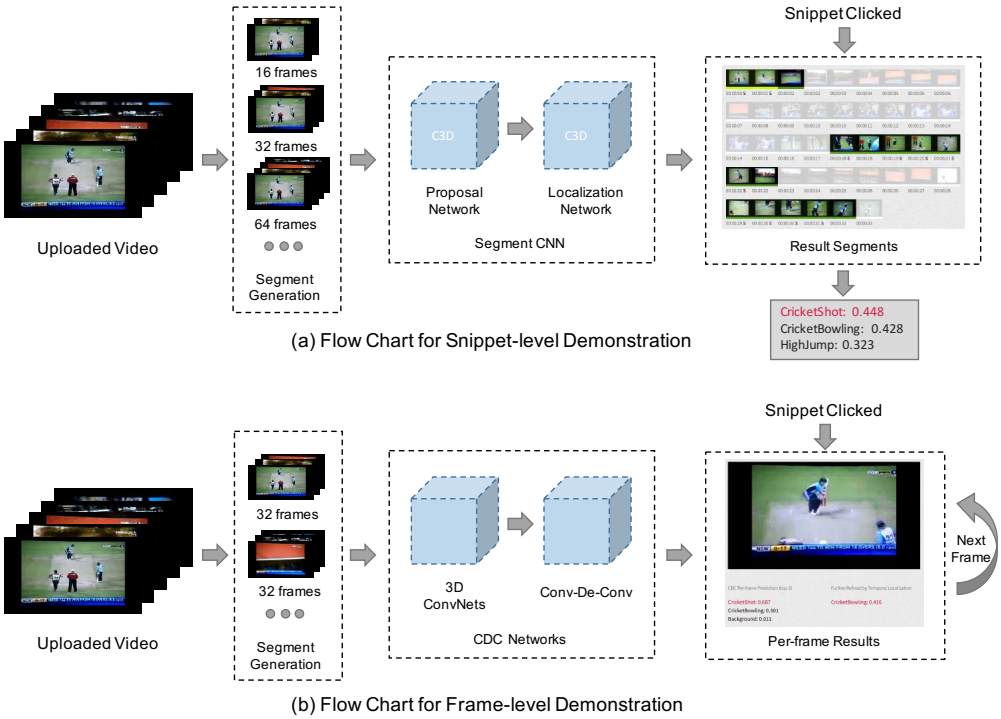


Figure 2: The flow charts of the two demonstration methods.

action detected. Therefore, using this method, users can clearly get the general content of the test video as well as the action locations.

The flow chart of the snippet-level demonstration is shown in Figure 2(a). As described in [5], the test video is first passed through the sliding windows and non-overlap segments with different lengths are generated. Then, the segments are put into the proposal network and localization network. After the post-processing, the results including action segments with action names, scores and start/end time are displayed as snippets. When a certain snippet is clicked, this snippet will play in the frame player, and the action names and scores will display below it.

2.3 Frame-level Demonstration

The frame-level method is designed for the frame-level TAL frameworks, such as CDC framework. In this method, the snippets are only used to navigate the time line of the test video, whereas the results of each frame will display synchronously with the frame.

The flow chart of the frame-level demonstration is shown in Figure 2(b). As described in [4], the test video is firstly split into non-overlap segments, and each segment has 32 frames. Then, these segments are put into the CDC networks and action detection results are obtained for each frame. When user clicks a snippet, the test video will start playing at this frame with detection results displaying below it.

2.4 Demonstration on THUMOS

In the demo, we also provide the pre-computed results for THUMOS 2014. For the THUMOS videos, our demo will not only show the test results, but also the ground truth of the video to give a intuitive feeling for users of the framework performance.

2.4.1 CDC Results for THUMOS. SCNN pre-computed results are similar with the results for uploaded videos. However, CDC pre-computed results are different because it also has segment-level results not only the frame-level results. Therefore, we use the idea of snippet-level method to show the performance of the segment-level results for CDC framework, which works well.

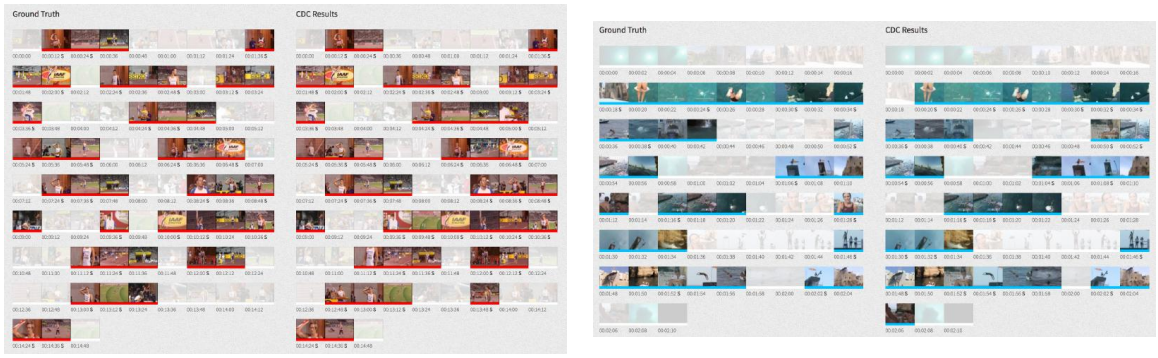
2.4.2 Some Example Results. There are some examples of loading THUMOS pre-computed results shown in Figure 3.

3 CONCLUSIONS

We developed a snippet-level and a frame-level demonstration for temporal action localization problem, which are used to show the results of SCNN and CDC frameworks respectively. These two demonstration methods can give users clear TAL results for the video they uploaded. In addition, all THUMOS 2014 test videos are available on the website, and users can see the TAL results of them and compare with the ground truth to evaluate the performance of SCNN and CDC frameworks.

REFERENCES

[1] Jake K Aggarwal and Michael S Ryoo. 2011. Human activity analysis: A review. *ACM Computing Surveys (CSUR)* 43, 3 (2011), 16.
[2] YG Jiang, J Liu, A Roshan Zamir, G Toderici, I Laptev, M Shah, and R Sukthankar. 2014. THUMOS challenge: Action recognition with a large number of classes. (2014).
[3] Ronald Poppe. 2010. A survey on vision-based human action recognition. *Image and vision computing* 28, 6 (2010), 976–990.



(a) CDC Snippet-level Results for JavelinThrow

(b) CDC Snippet-level Results for CliffDiving

Figure 3: Some sample results of CDC framework.

[4] Zheng Shou, Jonathan Chan, Alireza Zareian, Kazuyuki Miyazawa, and Shih-Fu Chang. 2017. CDC: Convolutional-De-Convolutional Networks for Precise Temporal Action Localization in Untrimmed Videos. In *CVPR*.

[5] Zheng Shou, Dongang Wang, and Shih-Fu Chang. 2016. Temporal Action Localization in Untrimmed Videos via Multi-stage CNNs. In *CVPR*.

[6] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*. 4489–4497.

[7] Daniel Weinland, Remi Ronfard, and Edmond Boyer. 2011. A survey of vision-based methods for action representation, segmentation and recognition. *Computer vision and image understanding* 115, 2 (2011), 224–241.