CSC 400 ITRG Mini-proposal

Temporal Action Localization

Songyang Zhang

September 2018

1. Project Introduction

Video Understanding, which involves the approaches in both computer vision area and nature language processing area, has attracted many researchers' attention in recent years. Many prior efforts focus on tasks like image recognition and object detection, which pave the way for some more challenging video tasks, like action recognition and temporal action localization.

Actually, temporal action localization is still unsolved, at least not as successful as object detection. The relationship between temporal action localization and object detection is quite close, many temporal localization methods are directly borrowed from object detection method, such as R-C3D [8], which have the similar two stage structure with Faster-RCNN [3]. However, compared to objects in spatial field, actions in temporal field usually have more redundant information and larger variations in size. Thus, many researchers attempt to adapt some successful object detection models into multi-scale models. Even so, the gain of performance is still unsatisfying. Since this problem is still unsolved, it is worthwhile for us to explore further.

2. Broader Impacts

Temporal action localization with different settings can applied to many different application areas. For example, Online Action Detection tries to detect action without knowing future frames, which is required in some online video services. Some researches concentrate on detecting and anticipating when the action start [4], which is used in applications that requires early response. Temporal action localization is demanded in many areas. For example, the natural language video retrieval [2], where sentence-based localization method is built on top of word-based video localization method, has raised in recent years. In abnormal event detection [6] or highlight

detection [9], some methods in this project can be directly applied to localizing abnormal event in surveillance videos or localizing highlights from web videos. Another example is spatial temporal action localization, which localize actions from both spatial and temporal domains. Both tasks require high quality action segments. Thus, it is necessary to solve temporal action localization first.

3. Project Description

Temporal action localization is the problem of temporally localizing actions in untrimmed video sequences. Specifically, given a long untrimmed video, the algorithm is required to detect all the action instances (including their action classes, start and end timestamps).

Due to the success in action recognition, many recent researches directly use the model pretrained on the recognition task to extract feature for temporal modeling. However, this may not be suitable for the detection task. The untrimmed video has a variety of background frames, which is unseen in pretrained model. A more decent way is to decompose each frame into actor's skeleton, the objects they might interacting with and the scene.

By explicitly modeling single actor's action, interactions between actors and interaction between actor and object, the localization precision might gain further improvement.

4. Research Plan

The research will be conducted in four phases:

a) first collecting a video dataset, which contains raw videos without dedicately clipping or other processing.
b) Then using faster-rcnn [3] to extract potential object in each frame on our video dataset.
c) Next using openpose [1, 5, 7] to detect actor's skeleton.
d) Finally, using the skeleton's and object's features to detect temporal localizations.

References

[1] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In CVPR, 2017.

[2] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In ICCV, 2017.

[3] Ross Girshick. Fast r-cnn. In Proceedings of the IEEE international conference on computer vision, pages 1440–1448, 2015.

[4] Zheng Shou, Junting Pan, Jonathan Chan, Kazuyuki Miyazawa, Hassan Mansour, Anthony Vetro, Xavier Gir´o i Nieto, and Shih-Fu Chang. Online detection of action start in untrimmed, streaming videos. In ECCV, 2018.

[5] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In CVPR, 2017.

[6] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. CoRR, abs/1801.04264, 2018.

[7] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In CVPR, 2016.

[8] Huijuan Xu, Abir Das, and Kate Saenko. R-c3d: region convolutional 3d network for temporal activity detection. In IEEE Int. Conf. on Computer Vision (ICCV), pages 5794–5803, 2017.

[9] Ting Yao, Tao Mei, and Yong Rui. Highlight detection with pairwise deep ranking for first-person video summarization. In CVPR, 2016.