Chapter 1

Related work

First approaches for action classification consisted of 2 steps a) compute complex handcrafted features from raw video frames such as SIFT, HOG, optical flow and b) train a classifier based on those features. These approaches made the choise of features a signifact factor for network's performance. That's because different action classes may appear dramatically different in terms of their appearences and motion patterns. Another problem was that most of those approaches take assumptions about the circumstances under which the video was taken because there was problems such as cluttered background, camera viewpoit variations etc.

Recent results in deep architectures and especially in image classification made us attempt to train CNN networks for the task of action classification and localization. As mentioned before, Action Localization can be seen as an extention of object detection problem, where the outputs are action tubes that consist of a sequence of bounding boxes. So, there are several approaches including an object-detector network for single frame action proposal and a classifier. [3] uses a 2-stream R-CNN [2] in order to generate action proposals for each frame. [8] NA DW TI KANEI.

[1] uses SSD

Some approaches include tracking [9]. Other approaches treat a video as a sequence of frames such as in [7] and in [6].

3d-2d pose

1.0.1 Our implementation

We propose a network similar to [6]. Our architecture is consisted by the following basic elements:

- One 3D Convolutional Network, which is used for feature extraction. In our implementaion we use a 3D Resnet network which is taken from [4] and it is based on ResNet CNNs for Image Classification [5].

- Tube Proposal Network for proposing action tubes (based on the idea presented in [6]).
- A classifier for classifying video tubes.

Bibliography

[1] Online real time multiple spatiotemporal action localisation and prediction on a single platform. *CoRR*, abs/1611.08563, 2016. Withdrawn.

[2] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *CoRR*, abs/1311.2524, 2013.

[3] Georgia Gkioxari and Jitendra Malik. Finding action tubes. *CoRR*, abs/1411.6031, 2014.

[4] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6546–6555, 2018.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[6] Rui Hou, Chen Chen, and Mubarak Shah. Tube convolutional neural network (T-CNN) for action detection in videos. *CoRR*, abs/1703.10664, 2017.

[7] Vicky Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, and Cordelia Schmid. Action tubelet detector for spatio-temporal action localization. *CoRR*, abs/1705.01861, 2017.

[8] Xiaojiang Peng and Cordelia Schmid. Multi-region two-stream R-CNN for action detection. In *ECCV - European Conference on Computer Vision*, volume 9908 of *Lecture Notes in Computer Science*, pages 744–759, Amsterdam, Netherlands, October 2016. Springer.

[9] Philippe Weinzaepfel, Zaïd Harchaoui, and Cordelia Schmid. Learning to track for spatio-temporal action localization. *CoRR*, abs/1506.01929, 2015.