Chapter 1

Related work

1.0.1 Action Recognition

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. A review techniques used until 2011 made by Aggarwal and Ryoo 2011.

Recent results in deep architectures and especially in image classification made us attempt to train CNN networks for the task of action classification. First significant attempt made by Karpathy et al. 2014. Simonyan and Zisserman 2014 and Feichtenhofer, Pinz, and Zisserman 2016 both added optical flow in order to achieve better results. On top of that, the increase in computing performance contributed to the design more complicated architectures including 3D Convolutions as presented in Ji et al. 2013 as done by Tran et al. 2014.

R(2+1) Tran et al. 2017 (Pending...More before 3D ResNet) Recent day 3D ResNet has been introduced by Hara, Kataoka, and Satoh 2018a

1.0.2 Action 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. The introduction of R-CNN (Girshick et al. 2013) achieve significant improvemet in the performance of Object Detection Networks. This architectures, firstly, proposes regions in the image which are likely to contain an object and then it

classify it using a SVM classifier. Inspired by this architecture, Gkioxari and Malik 2014 desing a 2-stream RCNN network in order to generate action proposals for each frame, one stream for frame level and one for optical flow. Then they connect them using viterbi connection algorithm. Weinzaepfel, Harchaoui, and Schmid 2015 extend this approach, performing frame-level proposals and using a tracker for connecting those proposals using both spatial and optical flow features. Also it performs temporal localization using a sliding window over the tracked tubes. Peng and Schmid 2016 used Faster R-CNN (Ren et al. 2015) instead of RCNN for frame-level proposals, and they use Viterbi algorithm for linking proposals, too. For temporal localization, they use a maximum subarray method.

Jain et al. 2014 introduces the tubelets.

Singh et al. 2017 uses SSD

Some approaches include tracking Weinzaepfel, Harchaoui, and Schmid 2015. Other approaches treat a video as a sequence of frames such as in Kalogeiton et al. 2017 and in Hou, Chen, and Shah 2017.

3d-2d pose

1.0.3 Our implementation

We propose a network similar to Hou, Chen, and Shah 2017. 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 Hara, Kataoka, and Satoh 2018b and it is based on ResNet CNNs for Image Classification He et al. 2015.
- Tube Proposal Network for proposing action tubes (based on the idea presented in Hou, Chen, and Shah 2017).
- A classifier for classifying video tubes.

Bibliography

Aggarwal, J.K. and M.S. Ryoo (Apr. 2011). "Human Activity Analysis: A Review". In: *ACM Comput. Surv.* 43.3, 16:1–16:43. ISSN: 0360-0300. DOI: 10.1145/1922649.1922653. URL: http://doi.acm.org/10.1145/1922649.1922653.

Girshick, Ross B. et al. (2013). "Rich feature hierarchies for accurate object detection and semantic segmentation". In: *CoRR* abs/1311.2524. arXiv: 1311.2524. URL: http://arxiv.org/abs/1311.2524.

Ji, S. et al. (2013). "3D Convolutional Neural Networks for Human Action Recognition". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.1, pp. 221–231. DOI: 10.1109/TPAMI.2012.59.

Gkioxari, Georgia and Jitendra Malik (2014). "Finding Action Tubes". In: *CoRR* abs/1411.6031. arXiv: 1411.6031. URL: http://arxiv.org/abs/1411.6031.

Jain, M. et al. (2014). "Action Localization with Tubelets from Motion". In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 740–747. DOI: 10.1109/CVPR.2014.100.

Karpathy, A. et al. (2014). "Large-Scale Video Classification with Convolutional Neural Networks". In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1725–1732. DOI: 10.1109/CVPR.2014.223.

Simonyan, Karen and Andrew Zisserman (2014). "Two-stream convolutional networks for action recognition in videos". In: *Advances in Neural Information Processing Systems*, pp. 568–576.

Tran, Du et al. (2014). "C3D: Generic Features for Video Analysis". In: *CoRR* abs/1412.0767. arXiv: 1412.0767. URL: http://arxiv.org/abs/1412.0767.

He, Kaiming et al. (2015). "Deep Residual Learning for Image Recognition". In: *CoRR* abs/1512.03385. arXiv: 1512.03385. URL: http://arxiv.org/abs/1512.03385.

Ren, Shaoqing et al. (2015). "Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks". In: *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'15. Montreal, Canada: MIT Press, pp. 91–99. URL: http://dl.acm.org/citation.cfm?id=2969239.2969250.

Weinzaepfel, Philippe, Zaïd Harchaoui, and Cordelia Schmid (2015). "Learning to track for spatio-temporal action localization". In: *CoRR* abs/1506.01929. arXiv: 1506.01929. URL: http://arxiv.org/abs/1506.01929.

Feichtenhofer, Christoph, Axel Pinz, and Andrew Zisserman (2016). "Convolutional Two-Stream Network Fusion for Video Action Recognition". In: *CoRR* abs/1604.06573. arXiv: 1604.06573. URL: http://arxiv.org/abs/1604.06573.

Peng, Xiaojiang and Cordelia Schmid (Oct. 2016). "Multi-region two-stream R-CNN for action detection". In: *ECCV - European Conference on Computer Vision*. Vol. 9908. Lecture Notes in Computer Science. Amsterdam, Netherlands: Springer, pp. 744–759. DOI: 10.1007/978-3-319-46493-0_45. URL: https://hal.inria.fr/hal-01349107.

Hou, Rui, Chen Chen, and Mubarak Shah (2017). "Tube Convolutional Neural Network (T-CNN) for Action Detection in Videos". In: *CoRR* abs/1703.10664. arXiv: 1703.10664. URL: http://arxiv.org/abs/1703.10664.

Kalogeiton, Vicky et al. (2017). "Action Tubelet Detector for Spatio-Temporal Action Localization". In: *CoRR* abs/1705.01861. arXiv: 1705.01861. URL: http://arxiv.org/abs/1705.01861.

Singh, Gurkirt et al. (2017). "Online Real time Multiple Spatiotemporal Action Localisation and Prediction". In:

Tran, Du et al. (2017). "A Closer Look at Spatiotemporal Convolutions for Action Recognition". In: *CoRR* abs/1711.11248. arXiv: 1711.11248. URL: http://arxiv.org/abs/1711.11248.

Hara, Kensho, Hirokatsu Kataoka, and Yutaka Satoh (2018a). "Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?" In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

— (2018b). "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)*, pp. 6546–6555.