Proposal for a PhD thesis

INRIA Sophia Antipolis, STARS group
2004, route des Lucioles, P93
06902 Sophia Antipolis Cedex-France

I Title

Action Detection for Untrimmed Videos based on Deep Neural Networks

II General objective

Temporal action detection in untrimmed videos (long video containing several actions) is an important task for monitoring patients, building robots for assisting and other healthcare applications. Although several approaches, including the Deep Convolutional Networks (CNNs), have significantly improved performance on action classification, they still struggle to achieve precise spatio-temporal action localization in untrimmed videos. Temporal action detection aims at not only recognizing the action category but also detecting the beginning and ending of an action instance. Most temporal action detection frameworks consist of two parts: action boundary proposition and action classification.

The first task, "action boundary proposal", consists in determining the temporal boundaries of each action instance. Existing work as [10, 14, 13, 8, 6] have low precision on this detection of temporal boundaries. These algorithms meet difficulties for detecting long complex actions (e.g. *cooking)*. Besides, they usually fail to detect the actions where the duration varies significantly, from a couple of seconds to few minutes. On the other hand, to obtain high localization accuracy, a large number of window scales and small sliding steps would be needed, which can lead to dramatically increased computational cost. Hence, we lack of an efficient and robust algorithm for localizing the actions.

The second task is "action classification" which is to classify accurately a video with action labels. Recently we have designed high performing model [1, 2], which can get more than 90% accuracy on several public datasets as NTU-RGB+D [9]. However, these models fail to achieve high performance in real life settings datasets. Errors come with handling real life challenges, such as high environment diversity, multi-view settings, low awareness of camera, high duration variation, etc. In addition, long action recognition with composite actions (e.g. making coffee *pour grain* and *pour water)* and fine-grained actions with different objects (e.g. *drinking from a cup* or *from a bottle*) are still unsolved tasks. Hence, we still need robust algorithms for action classification in real life settings.

The algorithm that we want to develop will be deployed in real life settings, to help senior people and their relatives to feel safer at home since video analytics intends to detect potentially dangerous situations and to report critical situations to caregivers.

To support this work, we have a full team of researchers specialized in human behaviors [15, 11, 3, 2], from experts in activity recognition, people detection and tracking, machine learning, up to medical doctors specialized in behavioral disorders. The STARS team has been working on analytics video understanding since 1994. The "SUP" ("Scene Understanding Platform") Platform developed in STARS, detects mobile objects, tracks their trajectory and recognizes related behaviors predefined by experts. This platform contains several techniques for the detection of people and the recognition of human postures and gestures of one person using conventional cameras. We have access to large cohorts of patients and can collect video datasets, dedicated to behavioral disorders, such as the ones induced by dementia. We have also large storage resources and a hefty GPU farm, from which 28 GPU nodes are dedicated to STARS team.

III Phd objective

In this work, we would like to go beyond Deep Learning by taking advantage of CNN based network for action classification and embedded them into a temporal action detection framework for action localization to address complex human daily living datasets.

The challenge is to design a method that can process an untrimmed video in both online and offline manner and so to detect automatically the beginning and end of the targeted actions. A typical system can include 2 sub-networks: generating temporal proposals and classifying proposed candidates. The former is to produce a set of class-agnostic temporal regions that potentially reflect actions of interest, while the latter is to determine whether each candidate actually corresponds to an action and what class it belongs to. CNNs, RNN could be used in this system.

The evaluation of proposed frameworks and models should be performed on public live videos and datasets which contain daily activities like AVA[5], THUMOS [4], PKU-MMD [7], DAHLIA [12] and Smarthome.

IV Prerequisites

Strong background in C++/Python programming languages,

Knowledge on the following topics is a plus:

Machine learning,
Deep Neural Networks frameworks,
Probabilistic Graphical Models,
Computer Vision, and
Optimization techniques (Stochastic gradient descent, Message-passing).

V Calendar

1st year:

Study the limitations of existing activity recognition and temporal detection algorithms. Depending on the targeted activities, data collection might need to be carried out. Propose an original algorithm that addresses current limitations on inference. Evaluate the proposed algorithm on benchmarking datasets. Write a paper.

2nd year:

Investigation of feasibility/appropriateness of the framework in practical situation. Propose an algorithm to address model learning task in semi-supervised settings, write a paper.

3rd year:

Optimize proposed algorithm for real-world scenarios. Write a paper and PhD Manuscript.

References

[1] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733. IEEE, 2017.

[2] S. Das, A. Chaudhary, F. Bremond, and M. Thonnat. Where to focus on for human action recognition? In *WACV 2019-IEEE Winter Conference on Applications of Computer Vision*, pages 1–10, 2019.

[3] S. Das, M. Thonnat, K. Sakhalkar, M. Koperski, F. Bremond, and G. Francesca. A new hybrid architecture for human activity recognition from rgb-d videos. In *International Conference on Multimedia Modeling*, pages 493–505. Springer, 2019.

[4] A. Gorban, H. Idrees, Y.-G. Jiang, A. R. Zamir, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes, 2015.

[5] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar, C. Schmid, and J. Malik. Ava: A video dataset of spatio-temporally localized atomic visual actions. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[6] M. Koperski. *Human Action Recognition in Videos with Local Representation*. Theses, Universite Cote d'Azur, Nov. 2017.

[7] C. Liu, Y. Hu, Y. Li, S. Song, and J. Liu. PKU-MMD: A large scale benchmark for skeleton-based human action understanding. In *Proceedings of the Workshop on Visual Analysis in Smart and Connected Communities*, pages 1–8. ACM, 2017.

[8] F. Negin, A. Goel, A. G. Abubakr, F. Bremond, and G. Francesca. Online detection of long-term daily living activities by weakly supervised recognition of sub-activities. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2018.

[9] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang. NTU RGB+D: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010– 1019, 2016.

[10] Z. Shou, D. Wang, and S.-F. Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1049–1058, 2016.

[11] U. Ujjwal, A. Dziri, B. Leroy, and F. Bremond. Late fusion of multiple convolutional layers for pedestrian detection. In *15th IEEE International Conference on Advanced Video and Signal-based Surveillance*, 2018.

[12] G. Vaquette, A. Orcesi, L. Lucat, and C. Achard. The daily home life activity dataset: a high semantic activity dataset for online recognition. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 497–504. IEEE, 2017.

[13] J. Yuan, B. Ni, X. Yang, and A. A. Kassim. Temporal action localization with pyramid of score distribution features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3093–3102, 2016.

[14] Y. Zhao, Y. Xiong, L. Wang, Z. Wu, X. Tang, and D. Lin. Temporal action detection with structured segment networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2914–2923, 2017.

[15] J. D. G. Zuniga, F. Bremond, et al. Residual transfer learning for multiple object tracking. In *International Conference on Advanced Video and Signal-based Surveillance (AVSS)*, 2018.