Proposal for a Phd Thesis

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

1 Title

Weakly-supervised activity detection in untrimmed videos using deep learning

2 General objective

Temporal activity detection in real world untrimmed videos (*i.e. long videos containing multiple activities*) plays a crucial role to assist elderly people in daily living, monitor patients in hospitals, guide children in indoor and outdoor scenarios. The goal of temporal activity detection is not only to recognize the activity category present in the video but also to detect precise starting and ending location of the activity instances. In real time scenarios, an untrimmed video may contain multiple long and short instances of activities which make the problem of detection and recognition challenging. The problem becomes more difficult when the activities are overlapped temporally with each other (*i.e multi-label instances*). Although, recent emergence of deep learning methods motivate the researchers to use them as a tool to solve the problem. However, these methods requires huge annotated data to learn the preciseness of the activities. As a matter of fact, acquiring large video data with dense temporal annotation is a labours task. Thus, to eradicate the waste of human labour and time development of weakly-supervised temporal activity detection algorithms are utmost need of the hour.

In weakly-supervised temporal activity detection algorithms only video-level labels are needed for training. However, with this information the algorithms not only predict exact locations of the activities but also classify the activities with accurate class label during testing. Existing methodology [2,8,11,12] on temporal activity detection suffers from the fact that they need strong temporal supervision during training, which is hard to obtain. Besides there are methods [6, 7] which follows weakly-supervised setting for activity detection, but they fails detect activities in long untrimmed video. However, for multi-label data the development of weakly-supervised algorithms still remains unexplored. Hence, we need to develop robust algorithms for temporal activity detection problem in real life settings.

To support this work, we have a full team of researchers specialized in human behaviors, 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.

3 Phd objective

In this work, we will take the benefits of CNN based networks [1, 9] used for action classification and human pose estimation, so that understanding of complex human behaviours can be addressed. Along with this, we will try develop novel cost function specifically suitable for weakly-supervised setting. Typical

framework can include CNNs for RGB feature extraction and pose estimation, LSTMs and TCNs for long range temporal modeling followed ranking cost functions to penalty the miss detection of the framework. As a major contribution we will also try to propose a new approach for the solution of weakly-supervised activity detection problem.

The evaluation of proposed frameworks and models should be performed on public datasets which contains activities of daily living. The publicly available datasets are THUMOS [4], PKU-MMD [5], Toyota Smarthome [3], DAHLIA [10].

4 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).

5 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] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[2] R. Dai, L. Minciullo, L. Garattoni, G. Francesca, and F. Bremond. Self-attention temporal convolutional network for long-term daily living activity detection. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–7, Sep. 2019.

[3] Srijan Das, Rui Dai, Michal Koperski, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. Toyota smarthome: Real-world activities of daily living. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[4] Haroon Idrees, Amir R Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The thumos challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding*, 155:1–23, 2017.

[5] Chunhui Liu, Yueyu Hu, Yanghao Li, Sijie Song, and Jiaying 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, 2017.

[6] 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, Nov 2018.

[7] Sujoy Paul, Sourya Roy, and Amit K Roy-Chowdhury. W-talc: Weakly-supervised temporal activity localization and classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 563–579, 2018.

[8] Zheng Shou, Dongang Wang, and Shih-Fu 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.

[9] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[10] Geoffrey Vaquette, Astrid Orcesi, Laurent Lucat, and Catherine 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.

[11] Jun Yuan, Bingbing Ni, Xiaokang Yang, and Ashraf 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.

[12] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2914–2923, 2017.