PhD Proposal

INRIA Sophia Antipolis, STARS group
2004, route des Lucioles, BP93
06902 Sophia Antipolis Cedex – France
http://www-sop.inria.fr/members/Francois.Bremond/

1. Title

Skeleton-Based Human Action Recognition in Real-World Videos

2. General objectives

Human action recognition approaches have led to significant contributions towards many current applications, such as video surveillance, video understanding, human-computer interaction and game control, making it an extremely active research field. There are two mainstream approaches, RGB-based action recognition using spatio-temporal deep convolution on the RGB videos and Skeleton-based approaches using the pose sequence as the input data. Recent skeleton-based methods draw increasing attention owing to their strong ability in summarizing human motion. As high level representations, skeletons have the merits of being robust to appearances, environments, and view-variant. However, in the real world videos full of occlusions and low-resolutions, there are two main challenges:

1) Human pose estimation in real-world videos. Many approaches like LCRNet++ [3], OpenPose [5], AlphaPose [4], etc. have attempted toward understanding realistic settings and becoming more robust to occlusion and they provide us with the pre-trained pose estimator so that we can extract skeleton data from real-world videos without expensive handcraft annotations.

2) Action recognition using pose. Deep-learning based approaches [1, 2] using RNNs or temporal CNNs are proposed owing to their high representation capacity but ignore the important semantic connectivity of the human body. Recent GCNs-based approaches [8, 7] construct spatial-temporal graphs and model the spatial relationships with GNNs directly, and these methods have seen significant performance boost, indicating the necessity of the semantic human skeleton for action recognition.

To support this work, STARS team works on automatic sequence video interpretation [14, 15, 16, 9] and in collaboration with Toyota Motor Europe, they have published Smarthome [6] dataset for daily living activity recognition which aims at detecting critical situations in the daily life of older people living at home alone. We believe that a system that is able to detect potentially dangerous situations will give peace of mind to frail older people as well as to their caregivers. 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 for the recognition of human postures and activities of one or several persons using 2D or 3D video cameras. In particular, there are 3 categories to recognize human activities:

1) Recognition engine using hand-crafted ontologies based on a priori knowledge (e.g. rules) predefined by users. This activity recognition engine is easily extendable and allows later integration of additional sensor information when needed.

2) Supervised learning methods based on positive/negative samples representative of the targeted activities which have to be specified by users. These methods are usually based on Bag-of-Words or fisher vectors or deep features computing a large variety of spatio-temporal descriptors.

3) Unsupervised (fully automated) learned methods based on clustering of frequent activity patterns on large datasets which can generate/discover new activity models.

However, there are many scientific challenges in recognizing human activities when dealing with real word scenes with dementia patients: lack of high-quality skeleton data, cluttered scenes, handling wrong and incomplete person segmentation, handling static and dynamic occlusions, low contrast objects, moving contextual objects (e.g. chairs)

3. PhD objective

In this work, we aim at designing the methods that can improve the performance of the skeleton-based supervised learning algorithm that automatically detects human activities of daily living. For this challenge, we can focus on many aspects, we can improve the performance of the pose estimation system to obtain higher quality skeleton data. We can also try to improve the structure of the skeleton-based Neural Network (GCNs, TCNs, CNNs, RNNs). Moreover, the performance can also be improved through data augmentation like view-invariant, temporal consistency action recognition.

On the other hand, we would like to have a generic (un)supervised pre-trained model that can be transferred to another dataset without retraining to reduce the cost of computations in applications.

The evaluation of proposed frameworks and models should be performed on public datasets which contains real-world activities like Toyota Smarthome [6], NTU-RGB+D [13], Kinetics [12], Charades [11], etc. There is a possibility of conducting an internship, before the PhD thesis.

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,
- Mathematic (Geometry, Graph theory, Optimization)
- Image processing and 3D Vision.

5. Schedule

1st year:

- Study of the limitations of existing skeleton-based action recognition algorithms on specified architectures. Proposing detailed research directions for proposing novel algorithms.
- Writing a paper.

2nd year:

- Designing new skeleton-based action recognition algorithms to address in particular:
    - Improving pose in the real-world videos even without pose annotations.
    - Improving the performance of AGCNs in the structure-level.
- Testing these algorithms on Smarthome dataset and Kinetics.
- Designing a novel unsupervised algorithm taking as input the 2D pose sequence and learning the view-invariant representation for downstream action recognition.
- Writing papers.

3rd year:

- Evaluating, improving and optimizing proposed action recognition algorithms by exploring the other research directions listed above. Writing a paper/journal.
- Oral presentation and Writing PhD manuscript.

6. Bibliography

1. Chao Li, Qiaoyong Zhong, Di Xie, and S. Pu. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. In IJCAI, 2018

2. Pengfei Zhang, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, and Nanning Zheng. View adaptive neural networks for high performance skeleton-based human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019

3. Grégory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. LCR-Net++: Multi-person 2D and 3D Pose Detection in Natural Images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019

4. 2019Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In ICCV, 2017

5. Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019

6. 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

7. Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In CVPR, 2019

8. S. Yan, Yuanjun Xiong, and D. Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. ArXiv, abs/1801.07455, 2018

9. Srijan Das, Saurav Sharma, Rui Dai, Francois Bremond, and Monique Thonnat. Vpn: Learning video-pose embedding for activities of daily living, In ECCV, 2020

10. S. Das, Arpit Chaudhary, F. Bremond, and M. Thonnat. Where to focus on for human action recognition? 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 71–80, 20

11. Gunnar A. Sigurdsson, Gul Varol, X. Wang, Ali Farhadi, I. Laptev, and A. Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In ECCV, 2016

12. J. Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. 2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR), pages 4724–4733, 2

13. 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

14. J. D. G. Zuniga, F. Bremond, et al. Residual transfer learning for multiple object tracking. In International Conference onAdvancedVideo and Signal-based Surveillance (AVSS), 2018.

15. 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.

16. Dai, Rui et al. “Self-Attention Temporal Convolutional Network for Long-Term Daily Living Activity Detection.” *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)* (2019): 1-7.