Proposal for a Phd Thesis

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

1. Title :

Video activity understanding to improve collaboration between humans and robots

2. Scientific context

There is an unprecedented economical and societal demand for robots that can assist humans in their industrial work. Unfortunately, the discrepancy between the expectations related to idealized versions of such assistive systems and the actual abilities of existing so-called collaborative robots is large. The limitations of existing systems include safety issues (collisions between robots and humans), inadequacy of the assistive tasks provided by the robots (e.g. providing the wrong tool), health hazards (e.g. repetitive stress injuries, muscular fatigue [1], wrong postures [2]) and cognitive stress induced by the adaptation to robots. These limitations are mostly due to the robot design process considering only a single operation performed by a unique human in an ideal environment (e.g. a laboratory), assuming that the human is an expert in robot interaction.

To go beyond these limitations, this PhD work proposes to study **human behaviours** while interacting with a collaborative robot using video monitoring. The goal is to provide useful feedback during the robot design process to the model representing the physical links and constraints between the human operator and the robot along different interaction modalities [3]. Such model will be elaborated through the PhD study proposed by Vincent Padois and David Daney from the Auctus Inria team: “A coupled view of the physical abilities of human-robot dyad for the online quantitative evaluation of assistance needs” conducted within the context of the LiChIE project.

In particular, this PhD work will provide a precise detection of complex human activities while operating with a robot in a real-world environment in order to determine:

- The profile of the human operator (e.g. novice, expert),
- The level of fatigue, stress of the human operator (e.g. the quality of the performed gesture),
- Whether there is a potential risk for collision/accident with the robot,
- What will be the most appropriate task that the robot could provide to the operator.

So, the goal of this work is not just to detect specific human activities in an untrimmed video stream, but to qualify the manner these activities are performed by the human operator.

3. Phd objective

Activity detection in real world untrimmed videos (i.e. long videos containing multiple activities) plays a crucial role to assist human operators in daily routines, in industrial work 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 makes 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 motivates the researchers to use them as a tool to solve the problem, these methods require huge amount of annotated data to learn the preciseness of the activities. As a matter of fact, acquiring large video data with dense temporal annotation is a tedious task. Thus, to eradicate the waste of human labour and time, the development of weakly-supervised temporal activity detection algorithms is the utmost need of the hour. In weakly-supervised temporal activity detection algorithms, only video-level labels (for instance, provided through the robot sensors) are needed for training. However, with this information the algorithms have not only to predict exact locations of the activities, but also to classify the activities with accurate class label during testing. Existing methodologies [5, 11, 14, 15] on temporal activity detection suffer from the fact that they need strong temporal supervision during training, which is hard to obtain. 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. Another challenge consists in qualifying the manner these activities are performed by the human operator (e.g. the quality of the posture or gesture while interacting with the collaborative robot) [9]. To support this work, we have a full team of researchers specialized in human behaviours, from experts in activity recognition, people detection and tracking, machine learning, up to medical doctors specialized in behavioural 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 behaviours 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 collection of video datasets, dedicated to behaviour understanding. We have also large storage resources and a hefty GPU farm, from which 35 GPU nodes are dedicated to STARS team.

In this work, we will take the benefits of CNN based networks [4, 13] 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 architecture specifically suitable for weakly-supervised settings. Typical framework can include CNNs for RGB feature extraction and pose estimation, LSTMs and TCNs for long range temporal modelling followed by ranking cost functions to penalize the miss detection of the framework. As a major contribution, we will propose a new approach for the activity detection problem in real-world industrial settings. The evaluation of proposed frameworks and models will be performed on public benchmark datasets which contain activities of human operators in daily routines in the context of industrial work. Several publicly available benchmark datasets are instructional video datasets, such as COIN [7], IkeaFA [10] and HowTo100M [13]. An experimental validation of the proposed action monitoring will be also conducted together with the Auctus Inria team within the context of the LiChIE project and more particularly considering the general applicative context of satellites' manufacturing. This experimental work will require a focus on the online estimation of the activity being performed by the operator as well as an estimation of its associated postural state.

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

The time frame could be adapted.

1st year: Study the limitations of existing activity detection algorithms and the literature in the domain of existing assistance modes in collaborative robotics. Depending on the targeted activities, data collection might need to be carried out in collaboration with the Inria team, Auctus. 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 situations in collaboration of the Auctus team. Propose an algorithm to address model learning task in semi-supervised settings, and to qualify the manner the targeted activities are performed by the human operator. Write a paper.

3rd year: Optimize proposed algorithm for real-world industrial scenarios. Write a paper and PhD Manuscript.

References:

See: http://www-sop.inria.fr/members/Francois.Bremond/topicsText/myPublications.html

[1] Luka Peternel, Nikos Tsagarakis, Darwin Caldwell, and Arash Ajoudani. Robot adaptation to human physical fatigue in human-robot co-manipulation. Autonomous Robots, 42(5):1011-1021, 2018.

[2] Wansoo Kim, Marta Lorenzini, Pietro Balatti, Phuong DH Nguyen, Ugo Pattacini, Vadim Tikhano, Luka Peternel, Claudio Fantacci, Lorenzo Natale, Giorgio Metta, et al. Adaptable workstations for human-robot collaboration: A reconfigurable framework for improving worker ergonomics and productivity. IEEE Robotics & Automation Magazine, 26(3):14-26, 2019.

[3] Yeshasvi Tirupachuri, Gabriele Nava, Claudia Latella, Diego Ferigo, Lorenzo Rapetti, Luca

Tagliapietra, Francesco Nori, and Daniele Pucci. Towards partner-aware humanoid robot control under physical interactions. In Proceedings of SAI Intelligent Systems Conference, pages 1073-1092. Springer, 2019.

[4] 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.

[5] R. Dai, L. Minciullo, L. Garattoni, G. Francesca, and F. Bremond. Self-attention temporal convolutional network for long-term daily living activity detection. In 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pages 1−7, Sep. 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] Tang, Y., Ding, D., Rao, Y., Zheng, Y., Zhang, D., Zhao, L., ... & Zhou, J. (2019). COIN: A large-scale dataset for comprehensive instructional video analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1207-1216).

[8] 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.

[9] F. Negin, P. Rodriguez, M. Koperski, A. Kerboua, J. Gonzalez, J. Bourgeois, E. Chapoulie, P. Robert and F. Bremond. PRAXIS: Towards Automatic Cognitive Assessment Using Gesture Recognition, in the Expert Systems With Applications journal, ESWA-D-17-04080R1, Volume 106, Pages 21-35, 15 September 2018.

[10] Human Pose Forecasting via Deep Markov Models. Sam Toyer, Anoop Cherian, Tengda Han, Stephen Gould (2017). DICTA 2017

[11] 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.

[12] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatio-temporal features with 3d convolutional networks. In the IEEE International Conference on Computer Vision (ICCV), December 2015.

[13] Miech, A., Zhukov, D., Alayrac, J. B., Tapaswi, M., Laptev, I., & Sivic, J. (2019). Howto100M: Learning a text-video embedding by watching hundred million narrated video clips. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2630-2640).

[14] 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.

[15] 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