Proposal for a PhD thesis

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

I. Title

Activity recognition in videos combining ontology based language and CNN networks

II. General objective

Several investigations have been carried out to model activities of daily living (ADLs) to monitor older adults at home. Most systems have been developed using either simple sensor data (wearable sensors, touch sensors, RFID tags) or camera information to recognize ADLs in a home environment. However, existing work has either focused on simple activities in real-life scenarios, or the recognition of more complex (in terms of visual variabilities) activities in hand-clipped videos with well-defined temporal boundaries. We still lack research on methods that can retrieve several instances of complex activity in a continuous video (multimodal) flow of data. Existing methods that perform in online scenarios that can reason about the temporal and composite relations that characterize complex activities generally cannot handle uncertainty and tend to underperform in real life scenarios. Moreover they have difficulties to distinguish similarly looking activities.
On the other hand, methods that can handle uncertainty tend to ignore the temporal and composite relations of activities and learn short-term activity models directly from pixel data. Hence, latter model cannot recognize tong-term or composed activities. For instance, Deep Convolutional Neural Network CNN algorithms have been applied with great success to images and short videos, related to monitoring applications such as People Detection and Posture, Gesture and Action Recognition algorithms: DeeperCut (http://pose.mpi-inf.mpg.de/) and Real-time Multi-Person 2D Pose Estimation using Part Affinity Fields (https://www.youtube.com/watch?v=pW6nZXeWlGM&feature=share). In addition, current state-of-the-art algorithms focus on some specific actions (with low intra class variation) like for instance “chopping”. Hence, more generic actions like “cooking” can mean either “chopping” or “mixing”. Current methods do not perform well on distinguishing similar looking activities, like laying down and falling down.

Typical situations that we would like to monitor are Eating and drinking (how much? how often?) or Cooking (detect behavior that might lead to dangerous situations or non completion of the task).

The system we want to develop will help senior people and their relatives to feel more comfortable at their home, since scene understanding intends to help at recognizing potentially dangerous situations and reporting to caregivers if necessary.

IV. PhD objective

In this work we would like to go beyond Deep Learning by taking advantage of CNN for pose estimation or short action detection and embedded them into an ontology based framework for long term activity recognition to address complex human behaviors. Typical pipeline can include CNNs for pose-estimation and body part classification depending on the action to detect. Short temporal aspect of the actions can be handled through HMM, RNN or LSTM. The objective of these 2 steps is to extract meaningful mid-level features that can be further processed thanks to an ontology based reasoning engine. The ontology will be provided by the user to let him/her describe the targeted activities to be recognized. A challenge will be to propose an approach to leverage the knowledge acquisition process, in both part CNN processing and ontology based reasoning.

The evaluation of proposed frameworks and models should be performed on public datasets which contains everyday activities like Cooking Composite, Cooking 2, Breakfast , and domain-specific datasets like CHU (Nice Hospital – RGBD), ICP and GAADRD datasets [Kuehne et al, 2014; Rohrbach et al, 2015; Crispim-Junior et al, 2016].

There is a possibility of conducting first an internship, before the PhD thesis.

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 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 situations
Propose an algorithm to address model learning task in semi-supervised settings
Write a paper
Write PhD manuscript.

3rd year:

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

VI. Bibliography:

- P.H. Robert, A. Konig, S. Andrieu, F. Bremond, I. Chemin, P.C. Chung, J.F. Dartigues , B. Dubois, G. Feutren, R. Guillemaud, P.A. Kenisberg, S. Nave, B. Vellas, F. Verhey, J. Yesavage and P. Mallea. Recommendations for ICT use in Alzheimer's Disease assessment: Monaco CTAD expert meeting, JNHA - The Journal of Nutrition, Health and Aging Ref. No.: JNHA-D-13-00016R1, 2013.

- G. Sacco, V. Joumier, N. Darmon, A. Dechamps, A. Derreumeaux, L. Lee, J. Piano, N. Bordone, A. Konig, B. Teboul, R. David, O. Guerin, F. Bremond and P.H. Robert, Detection of activities of daily living impairment in Alzheimer's disease and mild cognitive impairment using information and communication technology, Clinical Interventions in Aging Volume 2012:7 Pages 539 - 549 DOI: http://dx.doi.org/10.2147/CIA.S36297, Link to PubMed, December 2012.

- C.F. Crispim-Junior, K. Avgerinakis, V. Buso, G. Meditskos, A. Briassouli, J. Benois-Pineau, Y. Kompatsiaris and F. Bremond. Semantic Event Fusion of Different Visual Modality Concepts for Activity Recognition, Transactions on Pattern Analysis and Machine Intelligence - PAMI 2016.

- P. Bilinski and F. Bremond. Contextual Statistics of Space-Time Ordered Features for Human Action Recognition. The 9th IEEE International Conference On Advanced Video and Signal Based Surveillance (AVSS 12), Beijing on 18-21 September 2012.

- P. Bilinski, E. Corvee, S. Bak and F. Bremond. Relative Dense Tracklets for Human Action Recognition . The 10th IEEE International Conference on Automatic Face and Gesture Recognition , FG 2013, Shanghai, China, 22-26 April, 2013.

- P. Bilinski and F. Bremond. Statistics of Pairwise Cooccurring Local SpatioTemporal Features for Human Action Recognition. The 4th International Workshop on Video Event Categorization, Tagging and Retrieval (VECTaR 2012) , ECCV 2012 Workshop, Firenze, Italy, October 13, 2012.

- Guido-Tomas Pusiol. Event Learning based on Trajectory Clustering. PhD, University of Nice-Sophia Antipolis, 30th of May 2012.

- Wanqing Li, Zhengyou Zhang, Zicheng Liu. Action Recognition Based on A Bag of 3D Points. IEEE International Workshop on CVPR for Human Communicative Behavior Analysis (in conjunction with CVPR2010), San Francisco, CA, June, 2010.

- Jiang Wang, Zicheng Liu, Ying Wu, Junsong Yuan. Mining Actionlet Ensemble for Action Recognition with Depth Cameras. IEEE Conference on Computer Vision and Pattern Recognition (CVPR2012), Providence, Rhode Island, June 16-21, 2012.

- Robust 3D Action Recognition with Random Occupancy Patterns. Jiang Wang, Zicheng Liu, Jan Chorowski, Zhuoyuan Chen, Ying Wu. ECCV 2012, Firenze, Italy, October 13, 2012.

- Ivan Laptev, Marcin Marszałek, Cordelia Schmid, Benjamin Rozenfeld. Learning Realistic Human Actions from Movies. CVPR 2008, 24-26 June 2008, Anchorage, Alaska, USA.

- C.F. Crispim-Junior, M. Koperski, S. Cosar and F. Bremond. Semi-supervised understanding of complex activities from temporal concepts. In Proceedings of the 13th IEEE International Conference on Advanced Video and Signal-Based Surveillance, AVSS 2016, in Colorado Springs, Colorado, USA, 24-26 August, 2016b.

- Lev *et al*., RNN Fisher Vectors for Action Recognition and Image Annotation, ECCV, Amsterdam, Netherlands, October, 2016

- Matthew Richardson and Pedro Domingos. 2006. Markov logic networks. Mach. Learn. 62, 1-2 (February 2006), 107-136. DOI=http://dx.doi.org/10.1007/s10994-006-5833-1

- D. Nitti, T. De Laet, L. De Raedt. Probabilistic logic programming for hybrid relational domains. Machine Learning, 103:3, pp. 307 - 449, Springer, 2016.

- H. Kuehne, A. B. Arslan and T. Serre. The Language of Actions: Recovering the Syntax and Semantics of Goal-Directed Human Activities. CVPR, 2014

- Recognizing Fine-Grained and Composite Activities using Hand-Centric Features and Script Data, M. Rohrbach, A. Rohrbach, M. Regneri, S. Amin, M. Andriluka, M. Pinkal, B. Schiele, IJCV 2015

VIII. Contact:

carlos-fernando.crispim_junior@inria.fr
Francois.Bremond@inria.fr