Understanding Human Activities from Observation via Semantic Reasoning for Humanoid Robots

Karinne Ramirez-Amaro[1], Michael Beetz[2] and Gordon Cheng[1]

I. Introduction

One of the main purposes of humanoid robots is to improve the quality of life of elderly and/or disabled people by helping them in their everyday activities. In other words, robots need to interact with humans in a meaningful, flexible and adaptable manner, ideally by observing, recognizing and understanding the behavior of the humans. Understanding human activities is a common and difficult problem in the areas of Artificial Intelligence and Robotics. Since this involves the use of our cognitive capabilities, e.g. perception, reasoning, prediction, learning, planning, etc. In other words, we understand, *what* we are doing. Namely, we extract the semantics of the observed behavior. Then, the ideal goal is to transfer such capabilities to robots so that they can better learn from us.

Automatically segmenting and recognizing activities from videos is a challenging task, due to the execution of a similar activity could be performed in many different manners. For example, if I prepare a pancake in my kitchen, then I may follow a predefined pattern. But, if I prepare a pancake in my office's kitchen under time pressure, then I will follow another pattern even though I execute the same task. These patterns are sometimes defined by different parameters, e.g. different speeds of execution, the height of the pancake mix to pour over the stove, force used to open a bottle, after how much time do I need to flip the dough, etc., which increase the dimensionality of the problem.

Our recent work [1] presented a framework for enabling robots *on-line* segmentation and recognition of human activities from observations. Our framework combines the information from different noisy physical sensors via semantic reasoning to enable robots to segment and recognize human activities by understanding what it sees from videos, i.e. using *high-level* representations (see Fig. 1). Our framework has been demonstrated to be robust to segment *on-line* human motions (*move, not move* and *tool use*) and object properties (*ObjectActedOn* and *ObjectInHand*) from videos.

II. Extracting the semantics from Videos

We propose to split the complexity of the recognition in two parts. The first one will gather (perceive) information from the objects using a simple color-based technique, whereas the second part will handle the difficult problem of interpreting the perceived information into meaningful classes using our inference module.

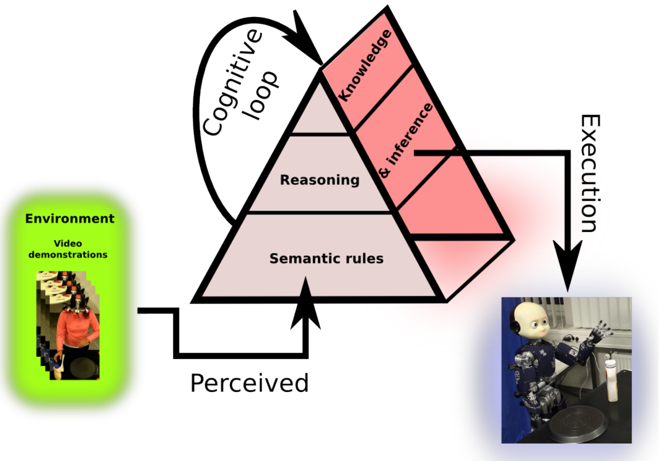

Fig. 1. This figure shows the main implemented modules of our system for the understanding of human everyday activities. First, we segment the human motions and object properties from videos. Then, we extract the semantics of the observed activity. Finally, we transfer the inferred activity to command the robot to achieve a similar activity.

The highest level of abstraction to be segmented from videos is the hand motions, into mainly three categories:

- *move*: The hand is moving, i.e. the hand velocity $\dot{x} > \varepsilon$
- *not move*: The hand stop its motion, i.e. $\dot{x} \rightarrow 0$
- *tool use*: Complex motion, the hand has a tool and it is acted on a second object, i.e. $o_h(t) = o_1$ and $o_a(t) = o_2$

Notice, that those kind of motions can be recognized in different scenarios, but they can not define an activity by themselves. Therefore, we need to add the object information, i.e. the motions together with the object properties have more meaning than separate entities. The properties that can be recognized from the videos are:

- *ObjectActedOn* (o_a): The hand is moving towards an object, i.e. $d(x_h, x_o) = \sqrt{\sum_{i=1}^{n}(x_{h_i} - x_{o_i})^2} \rightarrow 0$
- *ObjectInHand* (o_h): The object is in the hand, i.e. o_h is currently manipulated, i.e. $d(x_h, x_o) \approx 0$.

where $d(.)$ is the distance between the hand position (x_h) and the position of the detected object (x_o).

The output of this module determines the current state of the system (s), which is defined as the triplet $s = \{m, o_a, o_h\}$. The definition and some examples of the motions and object properties are explained in our earlier work [2].

In this paper, *the semantics of human behavior* refers to find a meaningful relationship between human motions

[1] Faculty of Electrical Engineering, Institute for Cognitive Systems, Technical University of Munich, Germany karinne.ramirez@tum.de and gordon@tum.de

[2] Institute for Artificial Intelligence, University of Bremen, Germany beetz@cs.uni-bremen.de

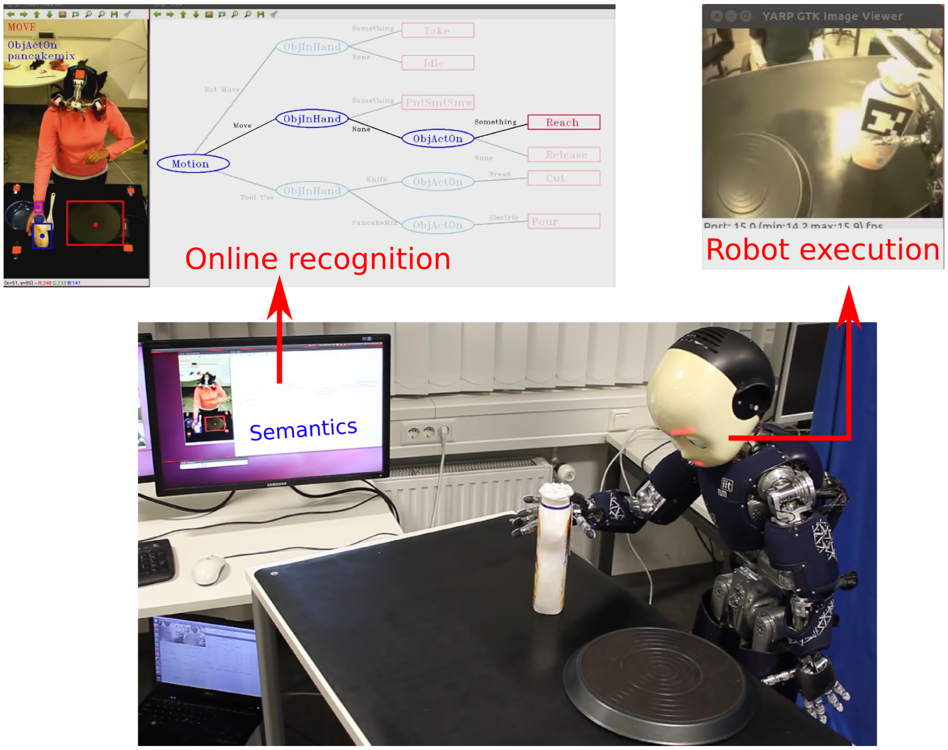

Fig. 2. Main steps made *on-line* by the robot to recognize and execute the observed human behavior. First the robot extracts the *relevant* information from a video. Then, it infers the human activity and finally the iCub executes a similar activity. https://www.youtube.com/watch?v=liYpFMCpyOE

and object properties in order to understand the activity performed by the human. In other words, *the semantics of human behavior* is used to interpret visual input to understand human activities. This has the advantage of transferring the extracted *meaning* into new scenarios.

This *semantic* module represents the core and most important part of our work. This module will interpret the visual data obtained from the perception module and process that information to infer the human intentions. In other words, it will be responsible of identifying and extracting the meaning of human motions by generating semantic rules, i.e. it will infer the *high-level* human activities, such as: *reach, take, pour, cut,* etc. The procedure to extract the semantics is further explained in our previous work [1], [2], [3].

III. Transferring the Models into Robots

Finally, we validate our *on-line* segmentation and recognition in a robotic system, in this case the iCub a 53 degrees of freedom humanoid robot. In other words, our framework considers *on-line* segmentation, recognition and the capability of learning new activities *on-demand* for the iCub. The robot is able to infer human activities from different scenarios with an overall accuracy of 85%, considering known and unknown activities. The above is possible even with a very simple hand and object recognition to segment the motions and object properties automatically. Noticeably, the communication between the perception and inference modules have to be instantaneous because these modules has to be implemented inside the control loop of the robot to have a smooth robot behavior (see Fig. 2).

IV. Conclusions

In this extended abstract we summarize our obtained results to automatically recognize human activities by extracting their semantic representations. Additionally, our system is adaptable, scalable and intuitive to new situations due to the re-usability of the learned rules. As a consequence, it can learn and identify new activities *on-demand*. Finally, our system is implemented inside the control loop of a humanoid robot and the obtained accuracy is still preserved around 85% accuracy. In other words, our system allows a more natural communication with artificial system such as robots.

Acknowledgments

The work leading to these results has received funding from the European Communitys Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 609206 and it was partly funded by the DFG cluster of excellence *Cognition for Technical systems CoTeSys*.

References

[1] K. Ramirez-Amaro, M. Beetz, and G. Cheng, “Automatic Segmentation and Recognition of Human Activities from Observation based on Semantic Reasoning ,” in *IEEE/RSJ IROS*. IEEE, Sept 2014.

[2] K. Ramirez-Amaro, E.-S. Kim, J. Kim, B.-T. Zhang, M. Beetz, and G. Cheng, “Enhancing Human Action Recognition through Spatio-temporal Feature Learning and Semantic Rules,” in *Humanoid Robots, 2013, 13th IEEE-RAS International Conference*, October 2013.

[3] K. Ramirez-Amaro, M. Beetz, and G. Cheng, “Extracting Semantic Rules from Human Observations.” in *ICRA workshop: Semantics, Identification and Control of Robot-Human-Environment Interaction.*, May 2013.