Bootstrapping Humanoid Robot Skills by Extracting Semantic Representations of Human-like Activities from Virtual Reality

Karinne Ramirez-Amaro[1], Tetsunari Inamura[2], Emmanuel Dean-León[1],
Michael Beetz[3] and Gordon Cheng[1]

***Abstract*— Advancements in Virtual Reality have enabled well-defined and consistent virtual environments that can capture complex scenarios, such as human everyday activities. Additionally, virtual simulators (such as SIGVerse) are designed to be user-friendly mechanisms between virtual robots/agents and real users allowing a better interaction. We envision such rich scenarios can be used to train robots to learn new behaviors specially in human everyday activities where a diverse variability can be found. In this paper, we present a multi-level framework that is capable to use different input sources such as cameras and virtual environments to understand and execute the demonstrated activities. Our presented framework first obtains the semantic models of human activities from cameras, which are later tested using the SIGVerse virtual simulator to show new complex activities (such as, *cleaning the table*) using a virtual robot. Our introduced framework is integrated on a real robot, i.e. an iCub, which is capable to process the signals from the virtual environment to then understand the activities performed by the observed robot. This was realized through the use of previous knowledge and experiences that the robot has learned from observing humans activities. Our results show that our framework was able to extract the meaning of the observed motions with 80% accuracy of recognition by obtaining the objects relationships given the current context via semantic representations to extract high-level understanding of those complex activities even when they represent different behaviors.**

I. Introduction

Enabling robots to learn new tasks, typically requires that humans demonstrate the desire task several times [1]. The observed motions should capture the human pose to further create the models that identifies the demonstrated task. However, this implies high costs due to the preparations of capturing new scenarios and those observations are limited to few tasks. Nevertheless, the problem of the acquisition of data can be (partially) solved using virtual environments when new scenarios and different conditions can be rapidly tested in a larger scale for more diverse scenarios than conventional means (see Fig. 1).

Virtual environments (VE) are human-computer interfaces in which the computer creates a sensory-immersing environment that interactively responds to and is controlled by the behavior of the user. For example, SIGVerse is a simulator environment, which combines dynamics, perception, and communication for synthetic approaches to investigate the genesis of social intelligence [2]. Using such simulator has several advantages, for instance fast and cheap set-up of new environments, different points of view of the analyzed scene, multi-user interaction, embodied interaction between virtual avatars and real user, etc. In other words, such VE are important tools specially when several human behaviors are investigated such as cooking, cleaning, etc. since they provide more complete and synchronize information about the executed task and the elements in the environment that greatly help in the understanding of human behaviors without the need of further expensive extra sensors.

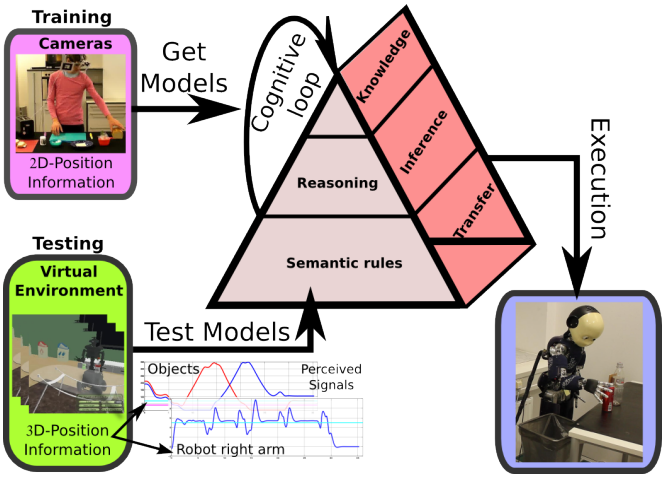

Fig. 1. Overview of our approach that is capable to re-use the learned models obtained from 2D cameras when using new input sources such as virtual environment signals as prosed in this work.

It is well known that the segmentation and recognition of human behaviors from observation is a difficult and challenging problem [3]. For this reason several alternatives have been proposed to gather the observed data, for instance using one static camera as presented in [4] or several external cameras, e.g. [5] and recently researches has been exploring egocentric cameras to analyze the human gaze information, e.g. [6], [7] to enhance the recognition of activities of daily life for robotic systems. However, those recordings are limited to the analysis of the obtained data and the acquisition of new tasks will require whole new set-ups, recruiting participants that will demonstrate the new tasks, accurate sensors located around the scenario, etc. which in long term represents a very costly and limited solution. Therefore, a better alternative is using virtual simulators that allow a long-term Human-Robot-Interaction due to its large

[1] Faculty of Electrical Engineering, Institute for Cognitive Systems, Technical University of Munich, Germany karinne.ramirez@tum.de and gordon@tum.de
[2] National Institute of Informatics, Japan. inamura@nii.ac.jp
[3] Institute for Artificial Intelligence, University of Bremen, Germany beetz@cs.uni-bremen.de

scale capabilities as proposed in [8].

In this paper, we propose a framework that is able to segment and recognize human behaviors based on previously learned experiences. The human activity recognition does not depend on the learned task and it is possible to be re-used in several and new scenarios using different input sensors, such as cameras and virtual scenarios. Fig. 1 depicts our proposed system, which first trains models to correctly identify the human behaviors while preparing a sandwich from real cameras. From this scenario, the semantic representations and reasoning engines are obtained using the observed cooking task. The challenging part is the transference of the obtained models into a new scenario where instead of observing a real human, we observe a virtual mobile robot, which is demonstrating a whole new task of *cleaning the table*. Finally, our system is fully implemented into a robotic platform that gathers the information of the VE and extracts the semantics of the demonstrated activity to understand the behavior of the virtual robot to execute a similar task in its real scenario.

In summary the main contributions of this paper are: a) we propose a multi-level framework that combines the information from different input sources such as 2D cameras and VE using our proposed semantic representations; b) the proposed framework is flexible and adaptable to new situations due to the re-usability of the learned semantics; c) we assess our framework using more complex tasks than the ones used on the training phase, thus demonstrating that our obtained models do not depend on the trained task; d) our presented framework is fully implemented on a humanoid robot that *imitates* the observed behavior from different input sources. The rest of this paper is organized as follows, section II presents the related work. Then, section III describes the technical details of the SIGVerse system. Afterward, section IV explains the steps executed on the virtual data. Then, section V presents the semantic representations method. Finally, section VI briefly expresses the obtained results followed by the conclusions.

II. Related work

The construction of hypothetical interaction models of humans should be designed to be large and preferable based on *big data* to make them more scalable and to achieve a more natural Human Robot-Interaction (HRI). Typically, research on HRI is done within a close laboratory, under very control scenarios, e.g. the light conditions are controlled, the location of the cameras, among other factors. However, that limits the exploration of more complex and difficult tasks normally analyzed in social and embodied interaction to build robust and general models about the studied interactions. This need was explicitly stated in the *Robohow*[1] project, whose goal is to enable robots to autonomously perform a large set of complex everyday manipulation tasks in real settings using websites, visual instructions and haptic demonstrations as primary information sources. However, integrating and combining those heterogeneous pieces is not a trivial task and it is still an unsolved problem.

Regarding the problem of human activity recognition several challenges have to be addressed, for instance: automatic segmentation of human motions [9], [10], identification of important features of the motion [11], definition of the importance of the object(s) to the task [12], as well as the definition of different levels of abstraction [13]. One of the main issues about those problem domains is that in order to translate the proposed methods from one task to solve a similar problem in another task is not straightforward. In other words, the recognition of human activities is still far from being an off-the-shelf technology [14].

Segmenting and recognizing human activities from demonstrations have been (partially) achieved using human poses mainly observed from external videos, e.g. using Conditional Random Fields (CRF) [15], Dynamic Time Warping [16], or by encoding the observed trajectories using Hidden Markov Models (HMMs) mimesis model [17]. However, the above techniques realize on the generation of trajectories which depend on the location of the objects, it means that if a different environment is being analyzed then the trajectories are altered completely, thus, new models have to be acquired for the classification, this implies that the proposed techniques need considerable time to finally *learn* a specific task [1]. Additionally, such techniques require a sophisticated visual-processing method to extract the human trajectories [18].

Recent studies focuses on determining the levels of abstraction to extract meaningful information from the produced task to obtain *what* and *why* certain task was recognized. Hierarchical approaches are capable to recognize high-level activities with more complex temporal structures [14]. Such approaches are suitable for a semantic-level analysis between humans and/or objects which can be modeled using object Affordances to anticipate/predict future activities [19], or using Graphical Models to learn functional object-categories [20], or Decision Trees to capture the relationships between motions and object properties [4]. For example, [21] suggests to use a library of OACs (Object-Action Complexes) to segment and recognize an action using the preconditions and effects of each sub-action which will enable a robot to reproduce the demonstrated activity. However, this system requires a robust perception system to correctly identify the object attributes which are obtained off-line. Then, based on the affordance principle, Aksoy et. al. [22] presented the called *Semantic Event Chain* (SEC), which determines the interactions between the hand and the objects, expressed in a *rule-character* form, which also depends on a precise vision system.

III. SIGVerse simulator

The SocioIntelliGenesis simulator (SIGVerse[2]) was principally developed for the RoboCup@Home simulation challenge [2]. SIGVerse enables a better and straightforward

[1]http://www.robohow.eu

[2]http://www.sigverse.org

HRI experiments, since all agents either real or virtual are able to interact socially and physically. Additionally, users can arbitrary join virtual HRI experiments trough Internet to enhance the interaction.

SIGVerse has three main modules: a) dynamics are used to simulate the physical properties of the objects and agents; b) perception which provides the senses of vision, sound, force and touch to enhance the HRI; c) communication between the available services.

SIGVerse is a client/server system consisting of a Linux server and a Windows client application. The server is in charge of running the dynamics calculations and of controlling the behaviors of the robot and human avatars. Whereas the Windows client is used to access the user interface in real time. In this work the server system was implemented to access the environment information, specially to obtain the position of the objects in the scene and the robot encoders information that can be recorded during the executed task as shown in Fig. 2.

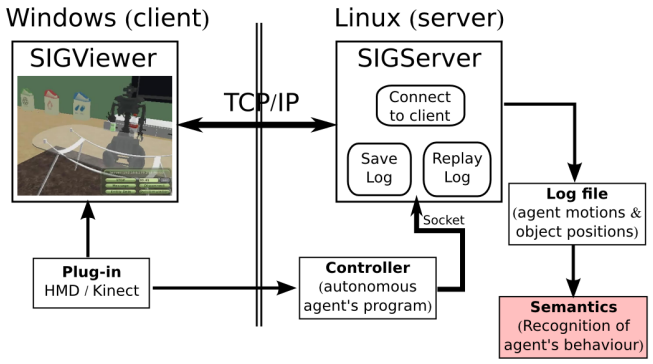

Fig. 2. Principal components of the SIGVerse software. Additionally, we can observe the communication between the server and client services. The output log file obtained from the virtual scenario is stored in a file that is used for the semantic system.

For our experiment, we choose the task of *cleaning up*, which is one of the challenge tasks according to the rulebook of the RoboCup@Home competition [8]. During this task the robot has to grasp a piece of trash targeted by the user and place it in a receptacle. Several problems are tested in this task and one of them is the understanding of the meaning of the instruction by speech recognition or by image processing of pictures captured with a camera. In this paper we will focus on the second problem.

The SIGVerse simulator is a state-of-the-art system that has recently gained attention at the RoboCup@Home competition in the simulation league due to its robust functionality. During this challenge users were able to control the robot trough a Joystick device or a Kinect device. In other words, the controller that determines the behavior of the robot used in this paper was written directly by random users. In this work the data used from SIGVerse contains the motions of a robot that is controlled by an autonomous controller module written by users.

IV. Extracting information from SIGVerse

In our previous work [4], we proposed a new method to recognize human activities based on semantic representations. This abstract method does not directly attempt to classify human activities, but rather, it infer the activities based on the observed human motions together with the information of the object of interest. To achieve this goal, we combine the visual information of the demonstrated task and the information of the human motions. First, we segment the continuous human motions into meaningful classes. Then, the second part handles the difficult problem of interpreting the perceived information into meaningful classes using our inference module. Three primitive human motions are segmented into mainly three categories:

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

Notice, that those kind of motions can be recognized in different scenarios and this method has been tested using as input: one camera [6] and multiple cameras [5]. However, these segmented motions 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 object properties that can be recognized are:

- *ObjectActedOn (o_a)*: The hand is moving towards an object, i.e. $d(x_h, x_o) = \sqrt{\sum_{i=1}^{n}(x_h - x_{o_i})^2} \to 0$
- *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 further explained in [5].

A. Results of the segmentation using virtual data

One of the main advantages of using VE is the fact that the location of agents and objects within the environment are known and can be acquired without any further perception system, which safes time when analyzing the data. However, our system previously presented in [4] only considered the information of 2D images. This means that we needed to adapt our system to also include 3D data. These changes are only reflected on the segmentation of the robot hand motions during the execution of the task *cleaning*, i.e. on the computation of the right end-effector velocities of the mobile virtual robot.

Since the obtained velocities of the virtual mobile robot end-effector presented some noise, we implemented a 2nd. order low-pass filter to smooth the obtained velocities. We choose the digital Butterworth filter with normalized cutoff

frequency Wn.

$$H(z) = \frac{b(1) + b(2)z^{-1} + ... + b(n+1)z^{-n}}{1 + a(2)z^{-1} + ... + a(n+1)z^{-n}} \quad (1)$$

where b and a are row vectors that contains the filtered coefficients in length $n+1$ where $n = 2$ and with coefficients in descending powers of z. This filter was also used to smooth the distance signal $d(x_h, x_o)$ before computing the object properties o_a and o_h.

Quantitatively the results of segmenting the motions (*move, not move, tool use*) of the virtual robot while executing the *cleaning* task are 77.83% accurate compare with the ground-truth[3]. Regarding the object properties for this virtual scenario are for the recognition of the *ObjectActedOn* property is 88.46% accurate and 90.5% when recognizing the *ObjectInHand* property.

V. Understanding Human Activities

In this work we propose two levels of abstraction: the *low-level*, which describes generalized actions such as: *move*, *not move* or *tool use*, and the *high-level* abstraction, which represents the basic human activities, such as: *reach, take, cut, release,* etc. Our technique uses the information from the *low-level* abstraction, to infer the *high-level* activities. This section briefly describes our method introduced in [4] to combine the observations obtained from the external cameras using semantic representations. In other words, this module interprets the visual data obtained from the perception module and process that information to infer the human intentions. This means that it receives as input information the hand motion segmentation (m) and the object properties (o_a or o_h).

In order to identify and extract the meaning of human motions, we used a decision tree to automatically generate the semantic rules that defines and explains the demonstrated human motions in a general manner. We used the C4.5 algorithm [23] and the Weka software to build our decision tree. We used as training data the information from 2D cameras for the sandwich-making scenario (see the *training* box from Fig. 1). Notice, that this scenario has high complexity due to the several sub-activities that it contains and different constraints.

Our proposed method consists of two steps to recognize human activities. For the first step, we used the information of the ground-truth data of the first subject for the scenario of making a sandwich. We split the data as follows: the first 60% of the trails was used for training and the rest 40% for testing. Then, we obtained the tree $T_{sandwich}$ shown in Fig. 3 which presents the same structured as the one previously presented in [4]. From this tree the following human *basic* activities can be inferred: *idle, take, release, reach, put* something somewhere and *granular*[4]. This learning process will capture the general information between the objects, motions and activities. It is important to highlight that a similar tree is obtained when the pancake data set is used for training as presented in [6].

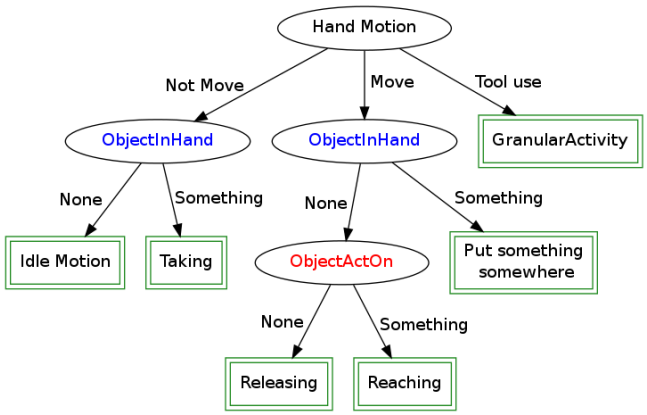

Fig. 3. This figure shows the tree obtained from the sandwich making scenario $(T_{sandwich})$.

From Fig. 3 we can observed that the activities: cut, sprinkle, etc. are inferred using the same rule:

$$if\ Hand(Tool\_use) \rightarrow Activity(GranularActivity) \quad (2)$$

This means that these activities need more information in order to be correctly classified. Those activities does not represent human basic activities, and we will call them *granular activities* and we used the second step of our proposed method to extend the obtained tree based on the current context to infer such activities in a similar manner as presented in [4].

A. Results of the human recognition

We tested the accuracy of the obtained tree $T_{sandwich}$ using the remaining 40% of the sandwich data set to validate the accuracy of the obtained rules. Then, given the input attributes $n_{sandwich\_test} = \{Move,\ Something,\ None\}$ we determine $c(n_{sandwich\_test})$. Afterward, the *state-value* pairs from the test data set $n_{sandwich\_test}$ are of the form $\langle n_{sandwich\_test}(t), ? \rangle$, where t represents the time $(frames)$. After that, the target value is determined for each state of the system $c(n_{sandwich\_test}(t))$. Finally, the obtained results show that $c(n_{sandwich\_test}(t))$ was correctly classified 92.57% of the instances using as input information manually labeled data, i.e., during the *off-line* recognition.

It is important to notice that the demonstrations used to train the semantic representations comes from 2D cameras and from a cooking activity. This implies that the human is always standing in the same location during the execution of the activity. Then, will it be possible to re-use the learned models depicted in Fig. 3 in a unknown scenario, where the demonstrator is moving constantly around the kitchen?.

In order to answer the above question, we tested the obtained tree from Fig. 3 with the new *cleaning* task using the information of the virtual environment, this means that now we have as input 3D information. Additionally, as can be observed in the attach video the agent that demonstrates

[3]The ground-truth data is obtained by manually segmenting the videos into hand motions, object properties and human activities.

[4]Granular activities define classes such as cut, pour, flip, etc. These activities are difficult to generalize because they depend on the context.

the behavior is a virtual robot which is moving around the kitchen, which makes the recognition even harder, specially when no previous training has been applied to this new situation.

From the obtained results we can notice that the activities *idle* and *release* are wrongly recognized by our system compared to the ground-truth. However, if we observe the obtained tree one in more detail (see Fig. 3), then we notice that in order to recognize the *idle* or *release* activities, we use the rules:

$$if\ Hand(NotMove)\ and\ ObjectInHand(None) \rightarrow Activity(\textbf{Idle}) \quad (3)$$

$$if\ Hand(Move)\ and\ ObjectInHand(None)\ and\ ObjectActedOn(None) \rightarrow Activity(\textbf{Realease}) \quad (4)$$

This indicates that in both activities the only difference is the motion of the hand, which is either *move* or *not move*. This is a very interesting phenomenon because it suggest that these two rules describe the same activity. This aspect did not pop up before since previously we have tested our system in different cooking scenarios, where the human is mostly standing in the same place, which is in front of a table. However, in this new scenario, we notice that the demonstrator, in this case the virtual robot, is moving around the kitchen, which stressed this situation.

Then, we considered the case when these two rules describe the same activity and we tested the sandwich making scenario again with this new assumption and the recognition improved from 92.57% to 95.43%. Additionally, we asked random participants to label the ground-truth data for the sandwich scenario and we noticed that they frequently misclassified the activities *idle* and *release*. Thus, demonstrating that this two rules are equivalent. Following a similar procedure, we compute the accuracy of recognition of our system into the new scenario and the quantitative results are 80% accurate compare to the ground-truth, which is a very high accuracy specially since no training was performed in this new scenario and the training and testing scenarios are very different.

VI. Transferring the models into humanoid robots

The experimental integration and validation of the acquired cognitive behavior into a humanoid robot are very important, essential and a challenging task. Therefore, as a final step, we validate our framework on a humanoid robot, the iCub a 53 degrees of freedom humanoid robot [24]. The implementation of our proposed framework within the control loop of the robot, follows a similar procedure as explained in our previous work [4]. However, we needed to include new procedures to adapt our code to the new scenario where more than one object is possible to be detected. Which means that several objects can have the same property at the same time. Then, to avoid that case, we improve our system specially, during the recognition of *ObjectActedOn* and *ObjectInHand* properties. This procedure is better explained in Algorithm 1.

Algorithm 1 Determine ObjActOn and ObjInHand.

Require: $distance[n]$: store the distance between the hand and objects detected (n).
$threshold\_OA$: ObjectActedOn threshold.
$threshold\_OH$: ObjectInHand threshold.
for $i = 0$ to N step 1 **do**
  if $distance[i] <= threshold\_OH$ **then**
    $OH\_vector.push\_back(i)$
  else
    if $distance[i] <= threshold\_OA$ **then**
      $OA\_vector.push\_back(i)$
    end if
  end if
end for
Find the value with lower distance from OA_vector to choose $ObjActOn$
for $index = 0$ to $OA\_vector.size()$ step 1 **do**
  if $distance[OA\_vector[index] < min$ **then**
    $ObjActOn = OA\_vector[index]$
    $min = distance[OA\_vector[index]]$
  end if
end for
A similar procedure is followed to find $ObjInHand$
return $ObjActOn, ObjInHand$

The results of the recognition of the human activities can be observed in Fig. 4. Our results also suggest that in order to execute a complex task such as *cleaning the table* it is possible to only recognize simple and *basic* human activities, such as: *reach, take, put, release/idle*. Another interesting result is the possibility to understand and execute complex activities such as *cleaning* without the use of the *granular* activities, which also demonstrates the robustness of our obtained semantic representations which are still valid (without any further training) under different scenarios and using different input sources such as cameras or VE.

Additionally, Fig. 4 depicts the integration between the *online* perception and semantic capabilities of our iCub robot to successfully recognize in *real time* human activities from different sources of information under different scenarios. In other words, we integrate and assess our system for different levels of complexity, i.e. first, we tested our obtained models using video inputs, later without further training we tested the same models under a new Virtual Environment scenario. Where the obtained results show that our framework is able to extract from the SIGVerse virtual simulator the meaning of the observed motions with 80% accuracy of recognition.

This indicates that our proposed system is designed in a way that allows different inputs without further modifications and without the need of further training to correctly extract the semantic of complex human activities as general as possible.

VII. Conclusions

In this paper we present our framework to bootstrap humanoid robot skills using virtual reality means via extraction of semantic representations for the recognition of human activities. Our semantic representations are obtained by segmenting *low-level* human motions, i.e, *move* or *not move* and two object properties, i.e. *ObjActedOn*, *ObjInHand*. We prove that our semantic rules captures the meaning of the human everyday activities, in a completely new scenario,

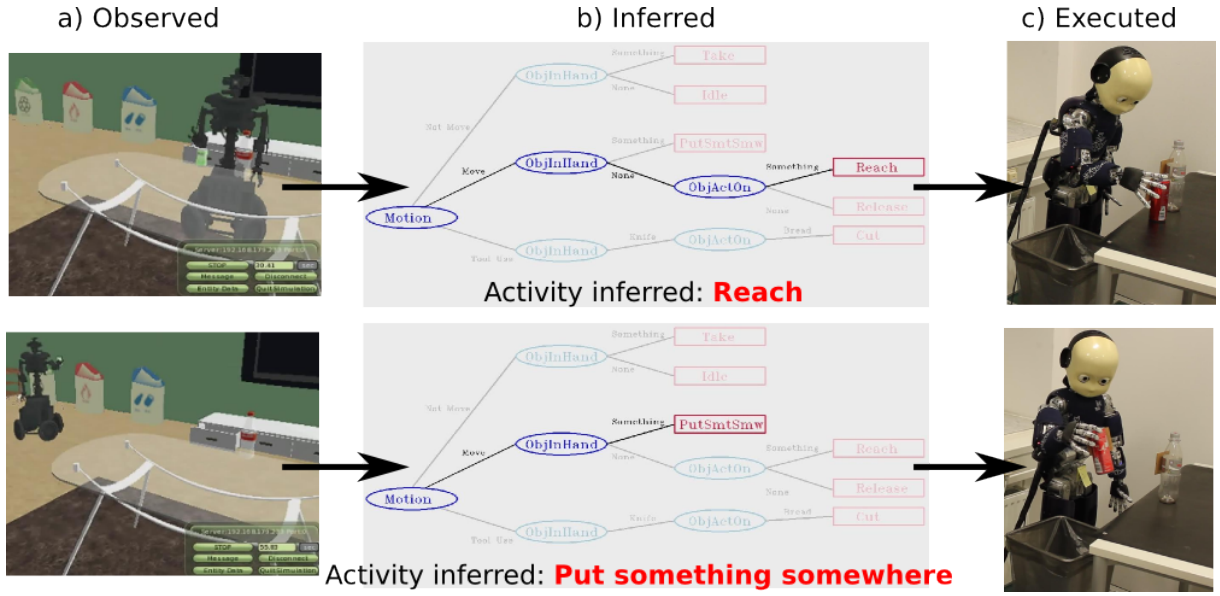

Fig. 4. First the robot observes the motions of the human from the external and the gaze videos, then it infers or learns the human activity and finally the iCub execute a similar activity.

i.e. *cleaning the table* without any further training with an accuracy of recognition around 80%.

ACKNOWLEDGMENTS

This work has received funding from the European Communitys Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 609206. This project was supported (in part) by the DFG cluster of excellence *Cognition for Technical systems CoTeSys*.

REFERENCES

[1] D. C. Bentivegna, C. G. Atkeson, and G. Cheng, "Learning Similar Tasks From Observation and Practice." in *IROS*. IEEE, 2006, pp. 2677–2683.

[2] T. Inamura, T. Shibata, H. Sena, T. Hashimoto, N. Kawai, T. Miyashita, Y. Sakurai, M. Shimizu, M. Otake, K. Hosoda, S. Umeda, K. Inui, and Y. Yoshikawa, "Simulator platform that enables social interaction simulation — SIGVerse: SocioIntelliGenesis simulator," in *IEEE/SICE Int. Symposium on System Integration (SII)*, Dec 2010, pp. 212–217.

[3] R. Poppe, "A survey on vision-based human action recognition." *Image Vision Comput.*, vol. 28, no. 6, pp. 976–990, 2010.

[4] 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.

[5] 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.

[6] K. Ramirez-Amaro, M. Beetz, and G. Cheng, "Extracting Semantic Rules from Human Observations." in *ICRA workshop: Semantics, Identification and Control of Robot-Human-Env. Int.*, May 2013.

[7] H. Pirsiavash and D. Ramanan, "Detecting activities of daily living in first-person camera views." in *CVPR*. IEEE, 2012, pp. 2847–2854.

[8] T. Inamura, J. T. C. Tan, K. Sugiura, T. Nagai, and H. Okada, "Development of RoboCup@Home Simulation towards Long-term Large Scale HRI," ser. RoboCup International Symposium, July 2013.

[9] R. Dillmann, "Teaching and learning of robot tasks via observation of human performance," *J. of Rob. & Auton. Sys.*, vol. 47, p. 2004, 2004.

[10] T. Taniguchi, K. Hamahata, and N. Iwahashi, "Unsupervised Segmentation of Human Motion Data Using a Sticky Hierarchical Dirichlet Process-Hidden Markov Model and Minimal Description Length-Based Chunking Method for Imitation Learning." *Advanced Robotics*, vol. 25, no. 17, pp. 2143–2172, 2011.

[11] V. Krüger, D. Herzog, S. Baby, A. Ude, and D. Kragic, "Learning Actions from Observations." *IEEE Robot. Automat. Mag.*, vol. 17, no. 2, pp. 30–43, 2010.

[12] M. Philipose, K. P. Fishkin, M. Perkowitz, D. J. Patterson, D. Fox, H. A. Kautz, and D. Hähnel, "Inferring Activities from Interactions with Objects." *IEEE Pervasive Comp.*, vol. 3, no. 4, pp. 50–57, 2004.

[13] K. Ikeuchi, M. Kawade, and T. Suehiro, "Toward assembly plan from observation - Task recognition with planar, curved and mechanical contacts," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, vol. 3, Jul 1993, pp. 2294–2301 vol.3.

[14] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: A review." *ACM Comput. Surv.*, vol. 43, no. 3, p. 16, 2011.

[15] M. Beetz, M. Tenorth, D. Jain, and J. Bandouch, "Towards Automated Models of Activities of Daily Life," *Tech. and Disability*, vol. 22, 2010.

[16] S. Albrecht, K. Ramirez-Amaro, F. Ruiz-Ugalde, D. Weikersdorfer, M. Leibold, M. Ulbrich, and M. Beetz, "Imitating human reaching motions using physically inspired optimization principles." in *Humanoids*. IEEE, 2011, pp. 602–607.

[17] W. Takano and Y. Nakamura, "Humanoid robot's autonomous acquisition of proto-symbols through motion segmentation," in *Humanoid Robots, 2006 6th IEEE-RAS International Conference on*. IEEE, 2006, pp. 425–431.

[18] P. Azad, A. Ude, R. Dillmann, and G. Cheng, "A full body human motion capture system using particle filtering and on-the-fly edge detection." in *Humanoids*. IEEE, 2004, pp. 941–959.

[19] H. S. Koppula and A. Saxena, "Anticipating Human Activities using Object Affordances for Reactive Robotic Response," in *Robotics: Science and Systems (RSS), 2013*, 2013.

[20] M. Sridhar, A. G. Cohn, and D. C. Hogg, "Learning Functional Object-Categories from a Relational Spatio-Temporal Representation." in *ECAI*, ser. Frontiers in Artificial Intelligence and Applications, M. Ghallab, C. D. Spyropoulos, N. Fakotakis, and N. M. Avouris, Eds., vol. 178. IOS Press, 2008, pp. 606–610.

[21] M. Wächter, S. Schulz, T. Asfour, E. Aksoy, F. Wörgötter, and R. Dillmann, "Action Sequence Reproduction based on Automatic Segmentation and Object-Action Complexes," in *IEEE/RAS International Conference on Humanoid Robots (Humanoids)*, October 2013.

[22] E. E. Aksoy, A. Abramov, J. Dörr, K. Ning, B. Dellen, and F. Wörgötter, "Learning the semantics of object-action relations by observation." *I. J. Robotic Res.*, vol. 30, no. 10, pp. 1229–1249, 2011.

[23] R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann Publishers, 1993.

[24] G. Metta, G. Sandini, D. Vernon, L. Natale, and F. Nori, "The iCub humanoid robot: an open platform for research in embodied cognition," in *PerMIS*, 2008, pp. 19–21.