Robust Semantic Representations for Inferring Human Co-manipulation Activities even with Different Demonstration Styles

Karinne Ramirez-Amaro, Emmanuel Dean-Leon, and Gordon Cheng

***Abstract*— In this work we present a novel method that generates compact semantic models for inferring human coordinated activities, including tasks that require the understanding of dual arms sequencing. These models are robust and invariant to observation from different executions styles of the same activity. Additionally, the obtained semantic representations are able to re-use the acquired knowledge to infer different types of activities. Furthermore, our method is capable to infer dual-arm co-manipulation activities and it considers the correct synchronization between the inferred activities to achieve the desired common goal. We propose a system that, rather than focusing on the different execution styles, extracts the meaning of the observed task by means of semantic representations. The proposed method is a hierarchical approach that first extracts the *relevant* information from the observations. Then, it infers the observed human activities based on the obtained semantic representations. After that, these inferred activities can be used to trigger motion primitives in a robot to execute the demonstrated task. In order to validate the portability of our system, we have evaluated our semantic-based method on two different humanoid platforms, the iCub robot and REEM-C robot. Demonstrating that our system is capable to correctly segment and infer *on-line* the observed activities with an average accuracy of 84.8%.**

I. Introduction

Enabling robots to learn new tasks typically requires that humans demonstrate the desire task several times [1]. This implies that the acquired models will greatly capture the execution style of the person demonstrating the desired activity. This however, could lead to a major problem for the generalization of the acquired model, since it will be almost impossible to teach robots all possible variations of one specific activity. Then, the ideal case is to enable robots with reasoning mechanisms to allow them to learn new activities by generating a model that interprets the demonstrated activities in a general manner, thus allowing the inclusion of different variations of the same activity.

Even when it is possible to observe stereotypical and predefined motion patterns from repetitive human movements [2], the execution of an activity or a similar activity can be performed in many different forms depending on the person, the place or the environment constraints. In other words, everybody has its own style to perform a desired activity. For example, Fig. 1 shows at least three different *real-life* demonstrations of the activity *cutting* the bread performed by random people. We can observe that the first participant is using a *common* style of holding the bread with his left hand and execute the *cutting* activity with his right hand,

Faculty of Electrical and Computer Engineering, Institute for Cognitive Systems, Technical University of Munich, Germany {karinne.ramirez, dean, gordon}@tum.de

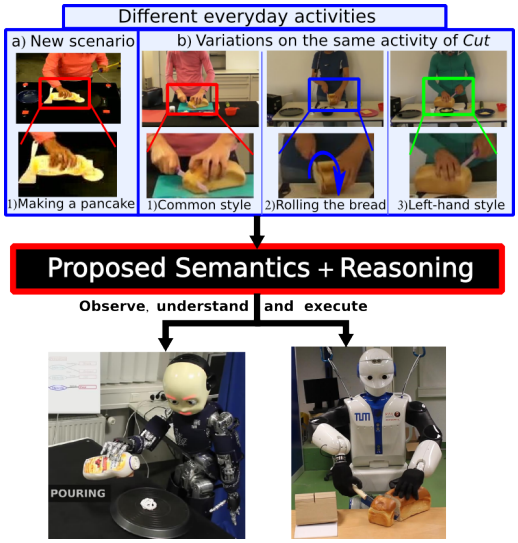


Fig. 1. Overview of our approach that is capable to deal with different demonstrations styles, since it extract the meaning of the observations.

Fig. 1.b.1. In a different form, the participant in the middle chooses to roll the bread with his left hand while *cutting* the bread with his right hand, Fig. 1.b.2. Finally, the third participant, which is left-handed, holds the bread with his right hand and *cuts* the bread with his left hand, Fig. 1.b.3. This indicates that we need to design a system that should handle all these different execution styles.

In our previous work [3], we presented a system which allows our iCub humanoid robot to extract the meaning of human activities using our proposed hierarchical approach to obtain semantic models. Later, we extended our system by including the activity recognition of both hands at the same time as presented in [4]. Even when our system is able to recognize activities of both hands, it was limited to only execute the activities of a single hand. Then, this paper enhances our current system in the following forms: a) we add robustness to variations on the demonstration styles for different activities; b) then, we include a method to tackle the problem of co-manipulation activities; c) finally, we improve the portability of our system and we tested it in two different humanoid platforms.

This paper is organized as follows, Section II presents the related work. Then, Section III describes the main modules of the presented system. Afterward, Section IV explains the steps to extract the *low-level* features. Then, Section V presents the semantic representations method. Section VI introduces the transference of the semantic models to

humanoid robots. Finally, section VII briefly expresses the obtained results followed by the conclusions.

II. Related work

Learning and understanding human activities can greatly improve the transference and generalization of the acquired knowledge among different robotic platforms. These robotic platforms have different embodiments and different cognitive capabilities [5], therefore the transference of trained models from one platform to another is not straight forward and typically the obtained models work fine only for the tested platforms [6]. However, if instead of learning *how* the motions are executed, we learn the meaning of such movements, then we can transfer these learned models to different situations and among different robotic platforms as proposed in this work.

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) [7], Dynamic Time Warping [8], or by encoding the observed trajectories using Hidden Markov Models (HMMs) mimesis model [9]. However, the above techniques realize on the generation of trajectories which depend on the location of objects and human postures, this means that if a different environment is being analyzed or its executed by a different demonstrator, 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 poses [10].

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 [11]. 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 [12], or using Graphical Models to learn functional object-categories [13], or Decision Trees to capture the relationships between motions and object properties [3]. For example, [14] 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. Based on the OACs principle, Yang et. al. [15] introduced a system that can *understand* actions based on their consequences, e.g. split or merge. Nevertheless, this technique greatly needs a robust active tracking and segmentation method to detect changes of the manipulated object, its appearance and its topological structure, i.e. the consequences of the action. Then, based on the affordance principle, Aksoy et. al. [16] 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. System modules description

Most of the recognition systems are designed to fit perfectly the studied task, however most of these systems can not easily scale toward new tasks or to allow different input sources [6]. This section presents the overall design and main components of our proposed framework. The main advantage of our system is its levels of abstraction that enhance its scalability and adaptability to new situations. For instance, our perception module permits the use of different input sources, such as: single videos [3], multiple videos [17] and virtual environments [18] which can bootstrap the learning process.

Our framework contains four main modules (see Fig. 2): 1) Perceive the relevant aspects of the demonstrations; 2) Generate or re-use semantic rules; 3) Infer the goal of the demonstrator; and finally, 4) Execute the inferred behavior by a robot.

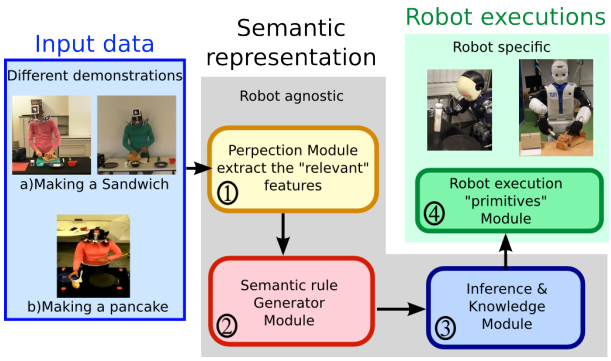


Fig. 2. Main modules of our proposed system

The *Perception Module*, enables robots to perceive different types of input information from different sources. This module extracts features obtained from the environment (input source), analyzes the features and generates meaningful information. Three primitive motions are segmented from the input data, i.e. *move*, *not move* and *tool use*. In addition, the information from the environment is also extracted, for example the objects in the scene and their properties, in our case we considered the following two properties: *ObjectActedOn* and *ObjectInHand*.

The *Semantic Module* represents the core of our system. It interprets the data obtained from the *Perception module* and processes that information to automatically infer and understand the observed human behaviors. It is responsible for identifying and extracting the meaning of human motions by generating semantic rules.

The *Inference module* includes knowledge-base to enhance the inference capability of our system. This is an important feature, in case of failure, the system uses its knowledge base to obtain an alternative and equivalent state of the system which contains instances of the desired class. With the knowledge-base we do not need to recompute the semantic rules every time we face a new situation.

The *Robot Primitives module* uses the inferred activity and enables an artificial system, i.e. allows a robot to reproduce the inferred action in order to achieve a similar goal as the one observed. This implies that given an activity, the robot needs to execute a skill plan and command the primitives from a library to generate a similar result.

Our system is based on a hierarchical approach, where different levels of abstraction are defined. This abstraction of the observation captures the *essence* of the activity, which means that our system is able to indicate which aspect of the demonstrator's activity should be executed in order to accomplish the inferred activity.

A. Levels of abstraction

Studies from the Psychology and Cognitive communities suggest that *goal-directed* activities exhibit recurrent inter-correlated patterns of activities [19]. Inspired by the above findings we propose our two levels of abstractions:

- The first one gathers (perceive) *low-level* features information from the environment, i,e. atomic hand motions such as: *move, not move* and *tool use*, as well as *basic* object properties, e.g. *ObjectActedOn* and *ObjectInHand*, (see Section IV);
- Whereas, the second part handles the difficult problem of interpreting the perceived information into meaningful classes using our proposed reasoning engine, i.e. the *high-level* human activities, such as: *reach, take, cut,* etc. (see Section V).

IV. Extracting low-level features from observations

In our previous work [3], we proposed to extract the following *low-level* features from observations, i.e three primitive human motions are segmented into one of the following categories:

- *move*: The hand is moving, i.e. $\dot{x} > \varepsilon$
- *not move*: The hand stops 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 $o_a(t) = bread$.

where $\dot{x}$ is the hand velocity. Notice, that those kind of *abstract* motions can be recognized in different scenarios, 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 as separate entities. The object properties that we define in this work 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 $d(\cdot,\cdot)$ 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 [4].

Since we expect noise on the perception of these *low-level* features, we implemented a 2nd. order low-pass filter to smooth the obtained velocities. We choose the digital Butterworth filter with normalized cutoff frequency.

V. Inferring high-level human activities

This module receives as input the state of the system (s), composed by the hand motion segmentation (m) and the object properties (o_a or o_h). Each of these perceived states populates the training sample (S), to infer the desired target function (c), which represents the *high-level* activities.

In this work, *the semantics of human behavior* refers to find meaningful relationships between human motions and object properties in order to understand the activity performed by the human. In order to achieve that, we use the C4.5 algorithm [20] to compute a decision tree (T), which learns the target function c by selecting the most useful attribute (A) that classifies as many training samples (S) as possible by using the information gain measure:

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{S} Entropy(S_v) \quad (1)$$

where *Values*(A) is the set of all possible values of the attribute A, and $S_v = s \in S | A(s) = v$ as a collection of samples for S.

Our proposed method consists of two steps to recognize human activities. The first one will extract the semantics of human basic activities, this means that target concept value is of the form:

$$Class\ c : \textit{ActivityRecognition} : S \to \{Reach,\ Take,\ Release,\ Put\_Something\_Somewhere,\ Idle,\ GranularActivity\} \quad (2)$$

where *GranularActivity* represents the set of activities that depend on the context. Therefore, to identify such kind of activities a second step is needed [4]. This step uses eq.(1) and the new target concept value:

$$Class\ c : \textit{ActivityRecognition} : S \to \{Cut,\ Spread,\ Sprinkle,\ Unwrapping,\ Dispose\} \quad (3)$$

A. Inferring parallel activities at the same time

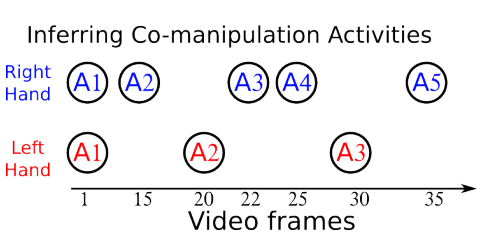
One of the major advantages of extracting semantic representations is its generalization capability. Even when the obtained tree (T) is trained to recognize the activities of a specific hand, it should be robust enough to also recognize similar activities performed by the other hand. This generalization capability is tested and exploited in our semantic-based framework, since the obtained Tree T (see Fig. 4) is robust when recognizing similar activities performed by both hands at the same time, without further changes in the algorithm.

B. Co-manipulation activities

One of the features of observing people from videos is to capture the natural way that people executes everyday activities in realistic environments, as the ones used in this work. However, the correct recognition of parallel activities sometimes is not enough, especially when dealing with collaborative activities, e.g. co-manipulation where two

hands work together to achieve a common goal. In this case, the correct activity recognition and their correct order of execution (synchronization) are both highly important to achieve the desired task.

For instance, in the case of *cutting the bread* task (see Fig. 1), it is important that the system first detects that the left hand moves the bread to the desired position before cutting the bread with the right hand. In order to cope with the co-manipulation problem, we proposed to build on demand a dependency table that will store the sequence where the observed activity is inferred as well as the hand executing such activity, as shown in Fig. 3.



Dependency table

Frames	Activity	Hand
1	A1	Left
1	A1	Right
15	A2	Right
20	A2	Left
22	A3	Right
25	A4	Right
30	A3	Left
35	A5	Right

Fig. 3. Proposed approach to deal with collaborative activities by automatically building a dependency table. Activities A1, A2, etc. represents the inferred activities, e.g. *reach*, *take*, *cut*, etc. Red labels represent the activities inferred from the left hand, and blue labels indicates the right hand activities.

From the obtained dependency table depicted in Fig. 3, our system is able to automatically generate execution sequences for the collaborative task from observations. For example that activity A_1 has to be executed by both hands at the same time. Then, activity A_2 of the right hand has to be executed before the activity A_2 of the left hand, in other words, the activity A_2 of the left hand, should not be triggered before A_2 of the right hand has finished. The system can identify when an activity has finished using the *Robot primitive module* of each hand. This module triggers different event flags for each activity: a) initialized, b) running or c) finished.

This analysis is very important since the activity recognition of each hand is made in parallel in our system and no information regarding the execution steps is known *a priori*. Additionally, when transferring our semantic-based system into robots, the correct execution sequence has to be considered to successfully achieve the desired collaborative task. This means that without the proposed dependency table, the correct execution sequence can not be guaranteed, thus leading to problems such as executing the *cutting* activity on the table instead of on the bread.

VI. Transferring the Models into Humanoid Robots

As a final step, we validate our framework on two humanoid robots[1], the iCub humanoid robot which has 53 degrees of freedom [21], and REEM-C a full-size biped humanoid robot[2]. To control the iCub robot, our system has been implemented in YARP as explained in our previous work [4]. However, for controlling the REEM-C robot we needed to transfer our modules into ROS. Then, to enable the recognition of both hands at the same time, we exploit the advantages provided by ROS using the *namespace* property for launching the same process multiple times in parallel.

It is important to highlight that the robotic system retrieves *on-line* the pose of the perceived hand and object(s) in the scene from the *perception module*, which triggers the *activity recognition module*, this means that the used videos have not been labeled in any way and the recognition is performed in *real-time*. This module sends the inferred activity to the *Robot Primitive module* to execute the proper primitive on the robot. For the REEM-C robot we have implemented the different skills for both arms in joint space. The *Perception* module obtains the 6D pose of the objects in the Cartesian space, and the desired joint position is computed using the Moveit![3] package in ROS. This desired position is the reference for the local joint position controller. For the *graps* primitive, a predefined grasp pose has been implemented, however more sophisticated controllers can be used such as the one proposed in [22].

VII. Results

For the obtained results, we use two data sets: *pancake* and *sandwich making*, which are publicly available[4]. The first one contains recordings of one human making a pancake several times. The second data set contains a more complex activity, which is making a sandwich performed by eight random subjects under two time conditions, i.e. *normal* and *fast*. This section presents the achieved results in two parts. First, we present the semantic representations obtained using our proposed method. Then, we present some results from the *on-line* implementation on the iCub and the REEM-C humanoids.

A. Results of human activities recognition

As mentioned in Section V, our proposed method consists of two steps to recognize human activities. For the first step, we use the information of the ground-truth[5] data of the first subject for the scenario of *making a sandwich*. We split the data as follows: the first 60% of the trails is used for training and the rest 40% for testing. Then, we obtained the tree $T_{sandwich}$ shown in the top part of Fig. 4 (magenta box) which infers the *basic* human activities defined in eq. (2).

Afterward, we apply the second step of our algorithm to further classify the *GranularActivities*, defined in eq. (3) and the extension of our tree is obtained as shown in the bottom part of the tree (T) depicted in Fig. 4.

We tested the accuracy of the obtained tree $T_{sandwich}$ using the remaining 40% of the sandwich data set to validate

[1]Another advantage of our proposed framework is its multi-robotic platform modality, in this case YARP (http://wiki.icub.org/yarp/) and ROS (http://wiki.ros.org/)

[2]http://pal-robotics.com/en/products/reem-c/

[3]http://moveit.ros.org

[4]http://web.ics.ei.tum.de/ karinne/Dataset/dataset.html

[5]The ground-truth was manually labeled by a person considered as an expert since this person received a training session.

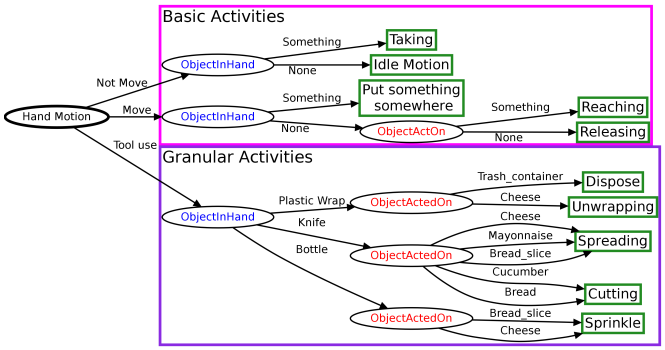


Fig. 4. This figure shows the tree obtained from the sandwich making scenario ($T_{sandwich}$).

the robustness 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.

B. Robot execution results

The next important step is to extend our system from *off-line* to *on-line* recognition using the Color-Based technique as presented in [4]. This implies that we use as input the data obtained from the automatic segmentation of human motions and object properties from the *perception module* in order to test the *on-line* recognition using the obtained tree $T_{sandwich}$. Compared with our previous work, we have improved the performance of our algorithm and the obtained results are presented in Table I.

TABLE I

OBTAINED RECOGNITION ACCURACY OF BOTH HANDS PER DATA SET.

Data Set	Accuracy of recognition (%)
Cut the bread (Common style-S01)	90.64
Cut the bread (rolling the bread-S04)	94.32
Cut the bread (left-hand style-S08)	75.41
Pour the pancake	84.42
Flip the pancake	79.18

First, we implemented our system in the iCub robot. In this case, we tested the data set of *pancake making* only for the right hand, where the activity *pouring* has not being included in the semantic model. However, from Fig. 4 it is possible to observe that the obtained taxonomy is preserved among the granular activities. This means that our system is able to learn new rules on demand using active learning methods. Thus, when the iCub recognizes that it does not know the new demonstrated activity, the system requires that the user provides a label for the new activity, in this case the label *pour*. Therefore, a new branch on the tree $T_{sandwich}$ is included to recognize this new activity in the future. Even when the system recognizes a new activity, this does not affect the inference accuracy, which in this case the average accuracy for both hands is 84.42%. A similar procedure is followed for the activity *flip* with an average recognition accuracy of 79.18%.

Another improvement of our system is its portability into different robotic platforms, taking into account a more complex scenario which includes parallel activities in collaboration. In this new scenario, the REEM-C robot must perform sequential activities with both hands. In this case, we are reusing the same semantic tree $T_{sandwich}$ to identify activities either for the right or the left hand, without affecting the recognition performance. Furthermore, the robot infers that the activity *reaching* is detected for both hands, therefore the robot executes both activities at the same time.

In addition, the system has been tested in a more complex case, depicted in Fig. 5. Here, the demonstrator performs the *cutting* activity in an odd-fashion, where she rolls the bread during the execution. Even with this odd-variation on the observation, our system is able to correctly infer the *cut* activity. Thus, demonstrating the robustness of our obtained semantic representations to the variations on the demonstrations of the same activity. Additionally, we observe that the left hand is executing the correctly recognized activity of *taking* the bread, preserving the correct sequence for the collaborative actions.

Notably, for these demonstrations of the same activity, we test the robustness of the obtained models, this means that no further training was performed to include the variance presented in these new observations. The average accuracy for the *on-line* segmentation and recognition of the overall activities for the scenarios shown in Table I is 84.8% for both hands. In our previous work [23], we concluded that the correct segmentation and recognition of human activities is not unique and greatly depends on the person interpreting the motions, especially when both hands are involved. The following link presents a video with more details about all these experimental results: http://web.ics.ei.tum.de/∼karinne/Videos/RamirezDChumanoids15.avi

VIII. CONCLUSIONS

One major problem of interpreting human everyday activities from observations is the fact that they greatly depend on the execution style of the person demonstrating the activities, leading to variations in the demonstrations. Although, each person has his/her own way of executing the same task, we (humans) have the capability to interpret all these different variations, since we abstract the meaning of the observations. Therefore, in this work we presented a framework that can deal with such problems in a robust manner with an *on-line* accuracy of recognition around 84.8%. Furthermore, our system can handle parallel activities in both cases: a) independent activities and b) in the more complex situation where the activities are sequential and collaborative, e.g. co-manipulation.

Our proposed method is able to automatically segment and recognize dual-arm human activities from observations using

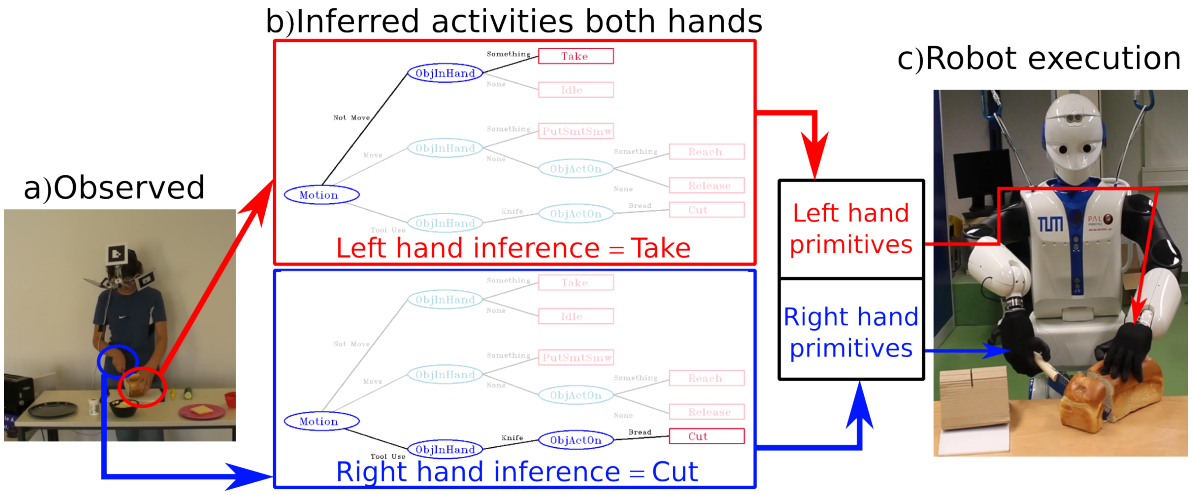


Fig. 5. The robot observes the motions of both hands of the human, then it infers for each hand the observed human activity, which in this case the right hand has been detected as *cutting*, while the left hand is *holding/taking* the bread. This means that the robot executes the activity *cut* using its right hand and *take* with its left hand.

semantic reasoning tools. Our presented system is flexible and adaptable to different variations of the demonstrated activities, as well as to new situations due to the re-usability of the learned rules, which allows the integration of new behaviors. Furthermore, our presented multilevel framework is applicable across different humanoid robots.

Acknowledgments

This work has received funding from the European Communitys Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 609206 and it was supported (in part) by DFG under the grant INST95/1284-1 FUGG; AOBJ:613263.

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] D. Wolpert and Z. Ghahramani, "Computational principles of movement neuroscience," *Nature Neuroscience Supplement*, vol. 3, pp. 1212–1217, 2000.

[3] 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 2014*. IEEE, Sept 2014.

[4] ——, "Understanding the intention of human activities through semantic perception: observation, understanding and execution on a humanoid robot." *Advanced Robotics*, vol. 29, no. 5, pp. 345–362, 2015.

[5] D. Vernon, G. Metta, and G. Sandini, "A Survey of Artificial Cognitive Systems: Implications for the Autonomous Development of Mental Capabilities in Computational Agents." *IEEE Trans. Evolutionary Computation*, vol. 11, no. 2, pp. 151–180, 2007.

[6] A. Billard, Y. Epars, S. Calinon, G. Cheng, and S. Schaal, "Discovering Optimal Imitation Strategies," *Robotics and Autonomous System*, vol. 47, no. 2–3, pp. 67–77, 2004.

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

[8] 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/RAS, 2011, pp. 602–607.

[9] W. Takano and Y. Nakamura, "Humanoid robot's autonomous acquisition of proto-symbols through motion segmentation," in *Humanoids, IEEE-RAS*. IEEE, 2006, pp. 425–431.

[10] 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/RAS*. IEEE, 2004, pp. 941–959.

[11] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: A review." *ACM Comput. Surv.*, vol. 43, no. 3, pp. 16:1–16:43, 2011.

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

[13] 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.

[14] 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 *Humanoids, IEEE/RAS*, 2013.

[15] Y. Yang, C. Fermüller, and Y. Aloimonos, "Detection of Manipulation Action Consequences (MAC)." in *CVPR*. IEEE, 2013, pp. 2563–2570.

[16] 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.

[17] 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 *Humanoids, IEEE/RAS*, October 2013.

[18] K. Ramirez-Amaro, T. Inamura, E. Dean-Leon, M. Beetz, and G. Cheng, "Bootstrapping Humanoid Robot Skills by Extracting Semantic Representations of Human-like Activities from Virtual Reality," in *Humanoids, IEEE/RAS*. IEEE, November 2014.

[19] R. Schank and R. Abelson, *Scripts, plans, goals, and understanding: an inquiry into human knowledge structures*. Lawrence Erlbaum Associates, 1977.

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

[21] 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: Performance Metrics for Intelligent Systems Workshop*, 2008, pp. 19–21.

[22] E. C. Dean-Leon, L. G. García-Valdovinos, V. Parra-Vega, and A. Espinosa-Romero, "Uncalibrated image-based position-force adaptive visual servoing for constrained robots under dynamic friction uncertainties." in *IROS*. IEEE, 2005, pp. 2983–2990.

[23] K. Ramirez-Amaro, M. Beetz, and G. Cheng, "Transferring skills to humanoid robots by extracting semantic representations from observations of human activities," *Artificial Intelligence. DOI:10.1016/j.artint.2015.08.009*, 2015.