Automated Models of Human Everyday Activity based on Game and Virtual Reality Technology

Andrei Haidu and Michael Beetz
{ahaidu, beetz}@cs.uni-bremen.de

***Abstract*— In this paper, we will describe AMEvA (Automated Models of Everyday Activities), a special-purpose knowledge acquisition, interpretation, and processing system for human everyday manipulation activity that can automatically (1) create and simulate virtual human living and working environments (such as kitchens and apartments) with a scope, extent, level of detail, physics, and close to photorealism that facilitates and promotes the natural and realistic execution of human everyday manipulation activities; (2) record human manipulation activities performed in the respective virtual reality environment as well as their effects on the environment and detect force-dynamic states and events; (3) decompose and segment the recorded activity data into meaningful motions and categorize the motions according to action models used in cognitive science; and (4) represent the interpreted activities symbolically in KnowRob[1] using a first-order time interval logic representation.**

I. Introduction

As we now have robotic agents that can accomplish mobile fetch&place tasks [2], [3], [4], the next challenge is how we can extend these capabilities into mastering human-scale manipulation tasks such as setting and cleaning the table, loading and unloading the dishwasher, or putting items back into cupboards. One of the biggest barriers in meeting these challenges will be the knowledge that the robotic agents have to be equipped with in order to accomplish these tasks successfully.

Consider for example the task of setting the table. If the robotic agent gets a task as underdetermined as "set the table" it needs a lot of knowledge to accomplish the task in the expected manner. It needs to know what is needed on the table, where the objects can be found, which objects can be used (you do not want to put dirty or broken plates on the table), how the objects should be arranged. All this depends on the meal that is to be served, whether it is casual or formal, whether a place is set for an adult or a small child, and other contexts.

The knowledge is not only needed to infer what has to be done but also for how it can be done. The robot needs to know where it has to position itself in order to pick up objects successfully, which hand to use, which grasp type to apply, where to position the fingers, how much grasp force to apply, how much lift force to apply, where to hold them, etc. Also, how to perform the fetch&place tasks efficiently, whether to use both hands, stack items, use a tray, leave cupboard doors open during table setting. In Figure 1 we illustrate a task execution by a human in a virtual environment and the visualization in openEASE[5] of the result of a query showing the world state and the trajectory of the spoon being transported by the right hand from the drawer to the tray.

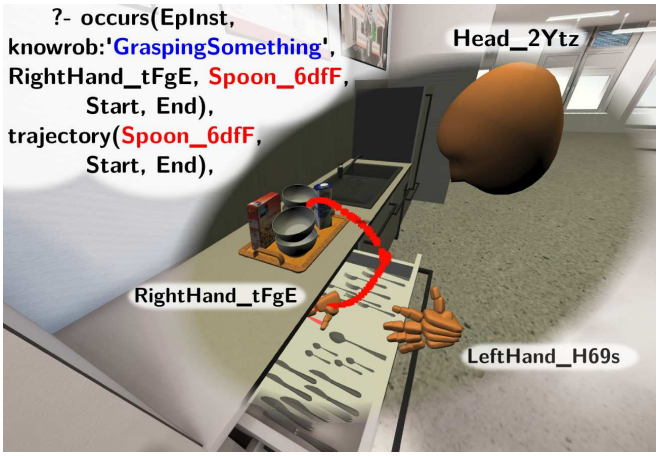

Fig. 1. Overlapping images of a fetch&place action execution in a virtual environment, illustrating: (1) the symbolic representation and visualization of the world state, including the trajectory of the manipulated item (center); and (2) the simulated kitchen environment (periphery)

Most of the manipulation tasks as we humans accomplish during the day are very knowledge intensive, even though it seems that we are not even consciously thinking about them. As Pratt has claimed in his article *"Is a Cambrian Explosion Coming for Robotics?"* [6]: *"The key problems in robot capability yet to be solved are those of generalizable knowledge representation and of cognition based on that representation."*. But before the knowledge can be represented and reasoned about it first has to be acquired. This is a particularly tricky task as the type of knowledge that is most critically needed is commonsense and naive physics knowledge, the knowledge that all humans have and apply without even being aware of it and often have difficulties in formulating it.

Modern technology, in particular games and virtual reality, gives us novel ways of acquiring commonsense and naive physics knowledge. Instead of engineering this knowledge [7], [8], [9] or crowd sourcing it [10], [11] we can set up tasks in games that require the commonsense and naive physics reasoning, as people to perform the tasks and mine the knowledge from the observed behavior.

In this paper we propose AMEvA (Automated Models of

The authors are with the Institute for Artificial Intelligence, University of Bremen, Germany.

This work was partially funded by Deutsche Forschungsgemeinschaft (DFG) through the Collaborative Research Center 1320, EASE.

Everyday Activities), a special-purpose knowledge acquisition, interpretation, and processing system for human everyday manipulation activity that can automatically

- create and simulate virtual human living and working environments (such as kitchens and apartments) with a scope, extent, level of detail and physics that facilitates and promotes the natural and realistic execution of human everyday manipulation activities;
- create a symbolic knowledge base from virtual reality environments that represents all objects in the environment, their parts and articulation models. This makes the system omniscient with respect to the environment. The knowledge base is extended with naive physics, commonsense, and background knowledge about the objects;
- record human manipulation activities performed in the respective virtual reality environment as well as their effects on the environment and detect force dynamic states and events;
- decompose and segment the recorded activity data into meaningful motions and categorize the motions according to action models used in cognitive science;
- represent the interpreted activities symbolically in KNOWROB using first-order time interval logic formulas linked to subsymbolic data streams;

We apply AMEVA to generalized fetch&place tasks (including organizing the kitchen, setting and cleaning the table, loading and unloading the dishwasher). The challenges are getting access to the relevant data structures of the game environment including objects, to the functional structure of the objects and their articulation models and to the force dynamic events that happen in the physics simulation of the environment.

We collect, manage, and provide public access to the observation data, models, and the symbolic representations of the activity episodes through the open and web-based robot knowledge service OPENEASE [5].

The remainder of this paper is organized as follows. In the following two sections we describe the functional view and the system architecture of AMEVA. In Section IV we introduce in detail the virtual environment. In Section V we present how the system observes, logs and recognizes activities. We thereafter show our collected experiments and results in Section VI.

II. FUNCTIONAL VIEW OF AMEVA

A key part of AMEVA is the automatic recognition of actions and events during task execution. For this we are using the model proposed in by Flannagan et al. [12], which has the concept, as illustrated in Figure 2, that actions are structured into motion phases, where phases have associated motion goals, which typically correspond to force dynamic events (distinctive events in the control apparatus). These motion phases have knowledge preconditions such as reaching motion, pregrasp pose, goal pose etc. By extracting regularities from the motion of people we can derive motion constraints and objective functions

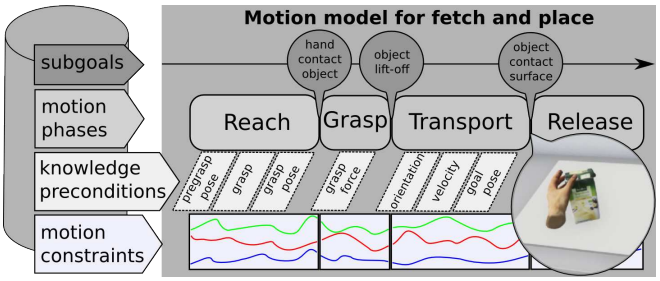

Fig. 2. Flannagan's fetch&place motion model

for constraint-based robot control. Figure 3 shows an example of automatically segmented action using the aforementioned model proposed by cognitive scientists. It segments the data according to the defined phases and can automatically extract all the knowledge preconditions such as the grasp and the pregrasp pose at the key frames, including trajectory data.

The purpose of AMEVA is to extend the commonsense knowledge that robots have, to be able to answer questions such as *"what are the arrangement of objects for table setting"*, *"where can I find a particular object for table setting"*, to learn motion constraints e.g. filled cups have to be held upright during transportation to avoid spilling. It not aiming for extracting robot specific information, hence the virtual worlds do not have to be configured to mimic the uncertainty that robots face in the real world.

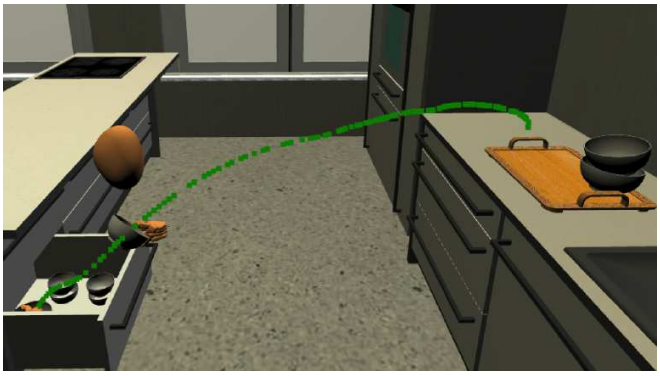

Fig. 3. Fetch&place activity recognition with scene reconstruction

III. SYSTEM ARCHITECTURE OF AMEVA

The system architecture of AMEVA is depicted in Figure 4. A detailed, realistic virtual model of a robot's working environment is created. In our case a kitchen environment, with cupboards, electrical devices with articulation and physical process models such as a simulated freezing process in the fridge. The environment is also equipped with objects of daily use such as plates, cups, milk boxes, knives, forks, etc. The environment has an integrated physics simulation that causes objects to fall down, move when they are pushed, and collide with other objects.

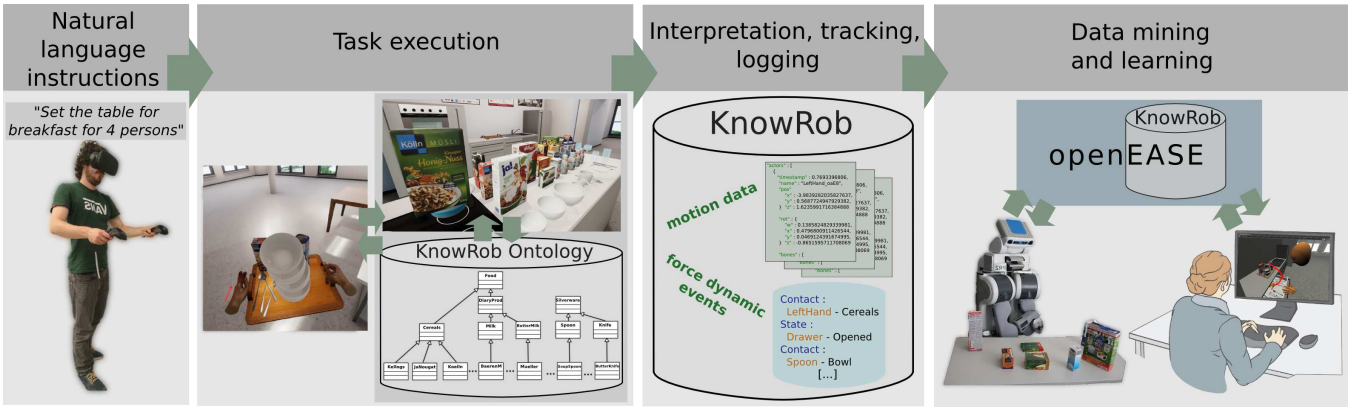

Fig. 4. System architecture of AMEvA: (1) users are asked to perform various tasks from underspecified instructions in a (2) virtual environment coupled with the robot's ontology. During execution (3) symbolic and subsymbolic data is automatically interpreted and logged to KnowRob. From the observed activities (4) robots can then learn generalized models of actions and motions.

To interact with the virtual world, the user uses an off-the-shelf virtual reality headset with hand tracking controllers. The tracked poses of the headset and the game controllers are then mapped onto the user's virtual head and hands. The mapping of the hands movements are done using force-based PD controllers, resulting in a more realistic interaction that in typical game environments. The virtual hands are fully rigged (e.g. every finger bone has a collision and is constrained by a joint) and collide with all the entities in the simulated world, thus the "harder" the user pushes/pulls against an object the larger is the force applied to it. The closing and opening of the hands are controlled by applying forces to the actuated finger joints. Considering that there is no finger tracking of the user's hands, these forces ar mapped from the analog button of the game controller. Since such force-based grasping is only possible in limited situations with well tuned physics engines, currently a fixation based grasping is used in the experiments.

The user is then instructed with a particular task such as "set the table for two people who will have coffee and cereals for breakfast". He/she will then walk through the virtual environment to the cupboard, open it in order to fetch the bowls for the cereals, he/she will then get spoons from the drawers and a milk carton from the fridge. Eventually, he/she will fetch the coffee cups, put them under the coffee machine and fill them.

All the objects in the virtual environment are coupled with the robot ontology of KnowRob. They all belong to a corresponding class in the ontology, this way a robot will be able to improve its knowledge acquisition skills by including background knowledge about the objects in his queries. For example it would know that the specific milk from the virtual world is of class type milk, which in turn is a dairy product, which has the property of being a perishable product. Subsequently it can discover that the fridge is a suitable place to store such products.

The natural setup for observing human-scale manipulation tasks in virtual environments can be used to acquire a variety of commonsense and naive physics knowledge that humans apply to accomplish their tasks successfully. For example, we can learn which objects humans believe to be necessary for having coffee and cereal for breakfast, where are these objects usually found, how should they be arranged on the table, how humans reach for, grasp, or hold them. The opportunities for commonsense and naive physics knowledge extraction from such natural everyday activities are diverse.

During the task execution: positions and orientations of all the entities in the virtual world are streamed to a database; objects states and their physical interactions are detected, interpreted and stored. In this interpretation the detection and categorization of force dynamic events is particularly important [13], [14]. For example, fetching and placing an object generates a sequence of force dynamic states where a hand *touches* the object to be fetched, the object *losing* contact to its supporting surface, then *making* contact with the supporting surface at the placing destination, followed by the hand *releasing* the object. These events are essential for understanding observed activities, because they can characterize and define action categorizes, thus allowing the segmentation of continuous motions into meaningful motion phases.

AMEvA then uses the motion data stream together with the time-synchronized force dynamic events in order to generate a hierarchical symbolic-subsymbolic activity representation. The symbolic part of this representation is stated in a first-order time interval logic [15]. In this formalism motion phases (and actions) are represented as *occurs(mp,[t1,t2])* asserting that motion phase *mp* occurs in the time interval starting at time instant *t1* and ending at time instant *t2*. If the motion phase is the object transfer phase of a fetch&place action then the force dynamic event losing contact to the supporting surface would occur at *t1* and the contact of the object with the supporting surface at the destination. By retrieving the respective pose data from the subsymbolic data stream AMEvA can retrieve the motion trajectory of the object and the hand

in the time interval *[t1,t2]*. Thus the result of the activity interpretation phase is a knowledge base that symbolically represents the observed activity and uses the symbolic part to retrieve subsymbolic data such as motions and poses of objects and body parts, as shown in Figure 1.

The last component of the AMEVA activity observation and interpretation system is action mining and machine learning. The purpose of this component is to learn generalized models of actions and motions that can be used by robots to fill knowledge gaps that are caused by underdetermined instructions. For example, the robot can learn the motion constraints for carrying open containers that are filled with substances.

In the remainder of the paper we will describe and discuss the components of the AMEVA system in more detail.

IV. ENVIRONMENT

Important factors for the acquisition of high quality action and motion data from virtual reality demonstrations is that humans performing the activities have the impression that the virtual reality environments look photorealistic, articulation models behave realistically, and the hand manipulation is intuitive and effortless.

Figure 5 shows the level of detail of the environment models. The depicted refrigerator is modeled through a detailed CAD model including a hierarchical object part model as well as a realistic articulation model of the door of the refrigerator. In addition, the presence of the dynamic light, in the form of the refrigerator light bulb, ensures realistic lightning conditions inside the refrigerator.

Besides the models of the devices and the pieces of furniture, the virtual kitchen environments also includes realistic models of objects of daily use, including mugs, bowls, spoons, forks, knives, cereal boxes, and the like. These object models allow us to create very realistic scenarios. For example, we can put all the objects of daily use on the kitchen counter and ask humans to put the items where they belong in order to learn the principles of how people organize their kitchens.

Fig. 5. Level of detail of the 3d models and textures in the virtual world

A. The virtual reality environment as a knowledge base

What sets observation and interpretation of human activities in virtual reality environments apart from the observation of real human actions through cameras or other observation means, are two key advantages. The first being that in virtual environments the activity interpretation algorithms have access to the simulation process and data structures. Therefore the algorithms have ground truth data on states and poses of objects as well as to the force dynamic states and events making activity interpretation easier, more accurate and fully automated. Second, we can use the data structures of the virtual environment in order to automatically create a symbolic knowledge base that contains all the relevant background knowledge about the objects in the environment. Knowing the purpose and the functionalities of the objects, supports the interpretation and the generalization of the observed activities, and the learning of generalized action and motion models.

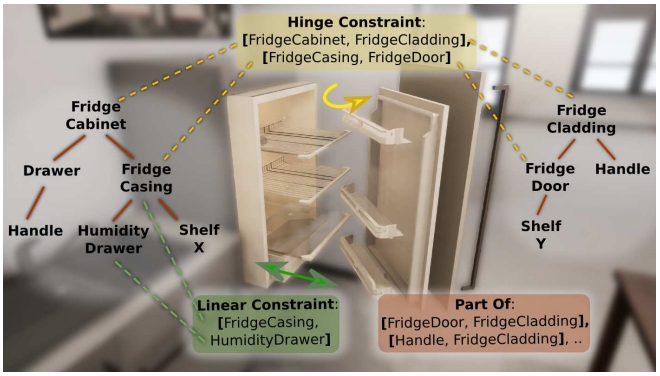

Fig. 6. Environment represented in the knowledge base, including hierarchies and articulations

When creating a new AMEVA environment, all the existing entity types (furniture, utensils etc.) need to be marked with their corresponding class type from the KNOWROB ontology. If the case, articulation properties or *part of* relations will be added as well. After this manual step, the system automatically assigns persistent unique identifiers for each entity and creates a semantic representation of the environment. This will assure that all future episodes will be commonly linked against the same representation, even if executed remotely or on multiple PCs. This allows the generated knowledge base to be constantly increased with new data.

Figure 6 shows the representation detail in the knowledge base of the environment, including parent-child hierarchies and articulations. We represent the hierarchies in a similar fashion to a semantic robot description language [16]. For the exemplified fridge cladding in the image, we have: a hinge joint between the door panel and the base; a linear joint between the bottom drawer and the base; and static fixations between the handles and the panel, respectively the drawer. Fluents, such as opening angles and poses of objects (or joints), are automatically updated and logged with their currently corresponding states (Opened,

Closed, HalfOpened, etc.) during execution.

V. Activity Observation

A. Representation of episodes

During the task execution AMEvA tracks the pose of each object, and their relevant parts, in the virtual world as well as the state of doors, drawers, knobs, etc. All tracking results are automatically optimized (filtering redundant data out) and logged into a noSQL database.

After each execution in the virtual environment, the resulting symbolic-subsymbolic activity representation data, coupled with the knowledge base can be loaded and visualized in openEASE. Figure 7 depicts the reconstruction of the virtual world at various key events from a logged episode. In the scenario the user was given the task so set the table for a 4 person breakfast. He was also instructed to use a tray in order to optimize his actions. The generated images illustrate the start of a fetch&place action, and highlight the trajectory of the grasped object during the action: grasping a bowl from the drawer and placing it on the tray; grasping the tray and placing it on the table; placing the milk from the tray on the table and taking the tray back to the kitchen island.

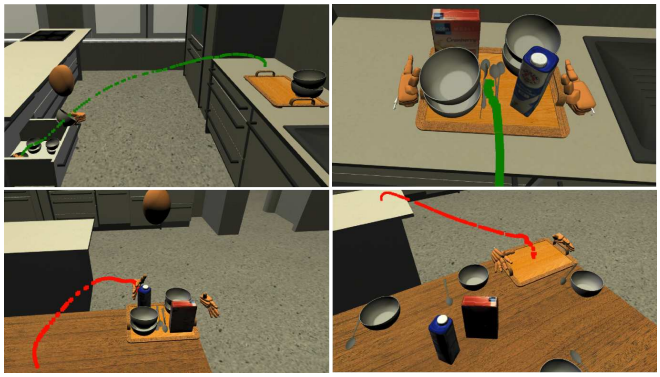

Fig. 7. Reconstruction and visualization of a recorded episode from the knowledge base

B. Recognizing force dynamic states and events

We propose to base action recognition on patterns of force interaction between entities. For example, we can characterize a fetch&place action through a sequence of force dynamic states and events. For fetch&place this sequence is (1) the object to be fetched supported by a supporting surface, (2) a hand touching the object, (3) the object being attached to the hand, (4) the hand not touching the object anymore, and the object being held by the supporting surface of the destination. Such a pattern is visualized Figure 2 from Section II.

Our hypothesis is that different actions can be characterized through their respective and distinct patterns of force dynamic interactions of objects and body parts and in particular hands. This view has been proposed in linguistics by Talmy who argued that the meaning of words can be effectively semantically categorized in terms of force dynamics and has later been adopted in action recognition and modeling in artificial intelligence [13].

The reason that the force dynamic characterization of actions is important for the action recognition in virtual environments is that the force dynamic states and events can be easily, reliably, and accurately detected by monitoring the physics simulation underlying the virtual reality environment. By computing the sequence of all relevant force dynamic states and events during an activity episode the interpretation algorithm can easily recognize and segment actions into the relevant motion phases.

In Figure 8 we can recognize the event of getting milk out of the fridge by comparing the various key frames in the executed action. The colored timelines from the image represent the automatically recorded events from the episode. We can see that the state of the fridge door changes while the hand is in contact with the door handle. After grasping the milk container, it is no longer in contact with the bottom shelf of the fridge, and it ends up on the table. Though, before ending up on the table we can observe that the free hand is again in contact with the door handle, and the state of the door is changing to closed.

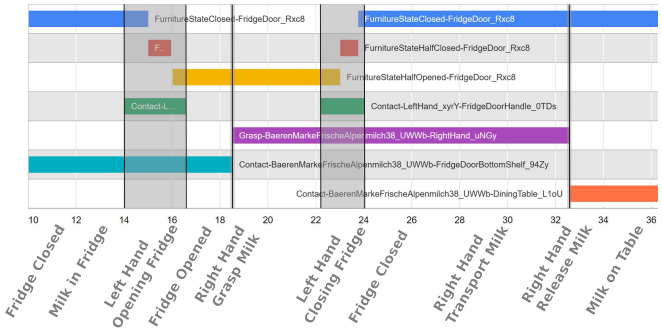

Fig. 8. Timelines showing the key frames of getting milk from the fridge

VI. Experiments

To showcase the capabilities of AMEvA we collected a set of 27 episodes where the users were given the task to set the table for a breakfast scenario with various variations:

- the number of persons to eat (1P, 2P, 4P);
- and, using 1 hand, 2 hands and 2 hands and a tray (1H, 2H, 2HT);

Being given the same task, or a similar one, to a robot, it could now use this data to optimize his steps during the task execution.

In Table I we have depicted the results of queries that check for the duration and the traveled distance during a correctly executed episode. The queries are created in a similar fashion as described in our previous work [17]. From the results we can notice that: time-wise it only starts to be useful to use a tray to carry items if the amount of persons to serve is at least 4 (or, the number of items to manipulate/carry is larger than the one used for this

particular scenario); traveled distance-wise, using a tray is the most advantageous; executing the task using only one hand is a definite disadvantage for both cases;

User	1P1H	1P2H	1P2HT	2P1H	2P2H	2P2HT	4P1H	4P2H	4P2HT
Duration (seconds)									
1	72.8	67.3	82.7	85.4	69.6	74.4	119.0	101.9	83.2
2	73.5	53.8	57.3	83.6	65.2	77.5	118.6	88.6	86.7
3	71.7	56.0	58.6	76.2	70.2	75.7	122.0	86.7	79.1
Avg	72.6	59.0	66.2	81.7	68.3	75.8	119,8	92.4	83.0
SD	0.74	5.91	11.6	3.98	2.22	1.27	1.51	6.76	3.10
Distance (meters)									
1	25.2	14.6	15.4	33.5	21.7	18.2	53.5	33.7	22.9
2	25.6	15.0	14.3	33.7	21.4	19.1	55.8	32.5	25.6
3	24.6	17.2	13.9	32.8	22.0	18.5	53.4	31.6	22.6
Avg	25.1	15.6	14.5	33.3	21.7	18.6	54.2	32.6	23.7
SD	0.41	1.30	0.63	0.38	0.24	0.37	1.10	0.86	1.34

TABLE I
EXPERIMENT RESULTS

Now assuming that a robot has prior knowledge about his skills and capabilities needed to execute this task (grasp objects, open drawers, carry items, etc.), it could now use this information to re-map the results for his own case. For example, if the robot is very slow at grasping objects, but doesn't have any issues with navigation/moving, then most probably using a tray would not be of its advantage since it will always include extra manipulation actions.

VII. RELATED WORK

Similarly to our approach, in [18] Bates et al. use a virtual reality environment as a viable way to collect semantic information about human behavior. They have set up a framework able to extract and reason on semantic data collected in real time. The user's known motions are continuously segmented and semantically classified by the system, while being capable to learn novel ones on demand. From the continuous observation of the users the system extracts the task space utilized by them in a form of a graph of all related activities. They use the KNOWROB ontology as well to store the classifications and the initial knowledge about the objects in the virtual world. In a similar fashion the virtual objects are tagged to their corresponding classes in the ontology. The focus of the paper was to recognize and learn new activities in complex virtual environments without prior training.

In [19] Fang et al. use a virtual environment from a robotic simulator to learn the relation between the physical effects a pouring action and the various variations in their execution style. Their proposed framework acquire and applies action knowledge from virtual user from naive user demonstrations in an interactive simulation environment under varying conditions. The authors argue that by collecting data from human users rather than a large set of automatically generated simulations, they can ignore a large proportion of the possible motion space. Using this data as a prior they can drastically reduce the computation time for learning parameters for the controller. They believe that using human demonstrations will help to construct a general framework capable of learning everyday manipulation skills.

VIII. CONCLUSION AND FUTURE WORK

In this paper we presented AMEVA (Automated Models of Everyday Activities), a special-purpose knowledge acquisition, interpretation and processing framework. The framework uses physics enabled and close to photorealistic virtual environments to promote a natural and realistic execution of human everyday activities. During execution the manipulation activities performed, as well as their effects on the environment, are recorded together with the detected force-dynamic states and events. The recorded activities are then decomposed and segmented into meaningful motions.

To showcase the capabilities of the framework we collected a set of 27 episodes where the users were given the task to set the table for a breakfast scenario with variations in the number of persons to serve and limiting the number of hands and tools to use. All the collected knowledge is symbolically represented in KNOWROB and available in the web-based knowledge service OPENEASE.

In our future research we plan to extend the action observation infrastructure to meal preparation tasks: we consider specific action verbs, including wiping, cutting, pouring and learn how to perform the respective actions for different objects, with different tools, and for different purposes. In the current state of implementation the objects of daily use are not modeled in depth. It is on our agenda to further detail the models so that milk cartons have a lid with an opening mechanism and that the containers are filled with virtual milk. With these deep models of objects of daily use we intend to enable high performance learning of everyday manipulation tasks.

We are working on introducing purely physics based grasping models, giving the users the possibility to change between various defined styles during runtime, thus ending up with a useful mapping of grasping styles to specific objects and scenarios.

We are advancing with the integration of full body tracking systems (Figure 9) using physics enabled movements. This would result in more natural movements, since kinematically impossible situations such as crossing the arms, switching hands, floating above objects etc. would be physically avoided. These movements are planned to be semantically mapped to a fully articulated human model represented in the knowledge base.

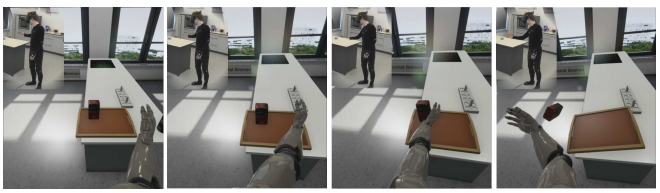

Fig. 9. Physics enabled full body tracking

REFERENCES

[1] M. Tenorth and M. Beetz, "KnowRob – A Knowledge Processing Infrastructure for Cognition-enabled Robots," *Int. Journal of Robotics Research*, vol. 32, no. 5, pp. 566 – 590, April 2013. [Online]. Available: http://ijr.sagepub.com/content/32/5/566.short

[2] N. Correll, K. E. Bekris, D. Berenson, O. Brock, A. Causo, K. Hauser, K. Okada, A. Rodriguez, J. M. Romano, and P. R. Wurman, "Lessons from the amazon picking challenge." *CoRR*, vol. abs/1601.05484, 2016. [Online]. Available: http://dblp.uni-trier.de/db/journals/corr/corr1601.html#CorrellBBBCHORR16

[3] L. P. Kaelbling and T. Lozano-Perez, "Hierarchical task and motion planning in the now," in *IEEE Conference on Robotics and Automation (ICRA)*, 2011, finalist, Best Manipulation Paper Award. [Online]. Available: http://people.csail.mit.edu/lpk/papers/hpnICRA11Final.pdf

[4] J. Winkler, F. Bálint-Benczédi, T. Fromm, C. A. Müller, N. Vaskevicius, A. Birk, and M. Beetz, "Knowledge-enabled robotic agents for shelf replenishment in cluttered retail environments," in *Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, Singapore, 2016.

[5] M. Beetz, M. Tenorth, and J. Winkler, "Open-EASE – a knowledge processing service for robots and robotics/ai researchers," in *IEEE International Conference on Robotics and Automation (ICRA)*, Seattle, Washington, USA, 2015, finalist for the Best Cognitive Robotics Paper Award.

[6] G. A. Pratt, "Is a cambrian explosion coming for robotics?" *Journal of Economic Perspectives*, vol. 29, no. 3, pp. 51–60, August 2015.

[7] D. B. Lenat and R. V. Guha, *Building Large Knowledge-Based Systems; Representation and Inference in the Cyc Project*, 1st ed. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1989.

[8] E. Davis, *Representations of Commonsense Knowledge*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1990.

[9] E. T. Mueller, *Commonsense Reasoning: An Event Calculus Based Approach*, 2nd ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2015.

[10] P. Singh, T. Lin, E. T. Mueller, G. Lim, T. Perkins, and W. L. Zhu, "Open mind common sense: Knowledge acquisition from the general public," in *Proceedings of the First International Conference on Ontologies, Databases, and Applications of Semantics for Large Scale Information Systems*, 2002, pp. 1223–1237.

[11] H. Liu and P. Singh, "Conceptnet — a practical commonsense reasoning tool-kit," *BT Technology Journal*, vol. 22, no. 4, pp. 211–226, Oct 2004. [Online]. Available: https://doi.org/10.1023/B:BTTJ.0000047600.45421.6d

[12] J. R. Flanagan, M. C. Bowman, and R. S. Johansson, "Control strategies in object manipulation tasks," *Curr. Opin. Neurobiol.*, vol. 16, no. 6, pp. 650–659, Dec 2006.

[13] L. Talmy, *Toward a Cognitive Semantics*, ser. Bradford book. MIT Press, 2000, no. v. 1. [Online]. Available: https://books.google.de/books?id=g7IoanNUNksC

[14] J. M. Siskind, "Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic," *CoRR*, vol. abs/1106.0256, 2011. [Online]. Available: http://arxiv.org/abs/1106.0256

[15] J. F. Allen, "Maintaining knowledge about temporal intervals," *Commun. ACM*, vol. 26, no. 11, pp. 832–843, Nov. 1983. [Online]. Available: http://doi.acm.org/10.1145/182.358434

[16] L. Kunze, T. Roehm, and M. Beetz, "Towards semantic robot description languages," in *IEEE International Conference on Robotics and Automation (ICRA)*, Shanghai, China, May, 9–13 2011, pp. 5589–5595.

[17] A. Haidu and M. Beetz, "Action recognition and interpretation from virtual demonstrations," in *International Conference on Intelligent Robots and Systems (IROS)*, Daejeon, South Korea, 2016. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7759439

[18] T. Bates, K. Ramirez-Amaro, T. Inamura, and G. Cheng, "On-line simultaneous learning and recognition of everyday activities from virtual reality performances," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2017)I*, IEEE, Ed. IEEE, 2017.

[19] Z. Fang, G. Bartels, and M. Beetz, "Learning models for constraint-based motion parameterization from interactive physics-based simulation," in *International Conference on Intelligent Robots and Systems (IROS)*, Daejeon, South Korea, 2016.