A Benchmarking Dataset for Automatic Symbolic Grounding from Virtual Demonstrations

Karinne Ramirez-Amaro[1], Constantin Uhde[1], Tamas Bates[2], and Gordon Cheng[1]

***Abstract*— This paper presents a new multi-purpose dataset obtained from virtual reality environments. The main goal of this dataset is to validate the scalability property of different learning methods, particularly semantic-based approaches for robotic applications. This dataset contains demonstrations from random participants performing three tasks, such as: *washing dishes*, *setting a table*, and *cleaning a room*. For each of the tasks different virtual scenarios were created, in this case *one kitchen*, *one dining room*, and *one living room*, respectively. The presented dataset contains information from 240 participants which capture different variations of the analyzed tasks. Thus, demonstrating its potential use for assessing the scalability and generalization properties of state-of-the-art learning approaches due to the high variance of the demonstrated tasks.**

I. Introduction

Recently, new emerging technologies such as virtual reality and wearable devices allow for capturing large amounts of natural human movements of users with ease. It is envisioned that the next generation of learning methods can take advantage of this and bootstrap the learning of new activities with these enhanced sensors by exploring large and diverse scenarios. The virtual reality systems, for example, open up the possibility to rapidly create new scenarios, as well as the fast analysis of objects and human movements without dealing with typical problems from perception methods such as occlusions when using camera recordings. Normally, one or several cameras are used to record human movements in household domains, in this case, the positions of the camera(s) will be fixed. Therefore, the perceived environment is limited to the positions of the camera(s). This limitation is more evident when the camera loses track of the observed object due to occlusions, e.g. one object is on top/in front of another, the participant executes the action in the opposite direction of the camera, etc. In this case, the data obtained from virtual scenarios gives enhanced information about the environment, for instance, 3D positions, 3D orientations, and velocities of every single object in the VR. For example, the VR systems can provide information about the objects located inside a drawer, even when these objects are occluded. However, obtaining the information of every single object in the scene brings new challenging problems, such as identifying the objects of interest. In general, if a learning algorithm is recognizing human activities, then objects such as walls, ceiling, and floor are not relevant for the recognition methods. Therefore, the learning method needs to filter out unnecessary data before it can create the desired models. To this aim, semantic-based methods have shown that they are able to deal with such problems in a reliable manner [1], [2], [3].

Fig. 1. Three different virtual scenarios: dining room, kitchen, and living room.

Semantic-based methods aim to find meaningful relationships between human motions and object properties in order to understand human activities. For instance, semantic-based approaches extract the meaning of the observed raw signals (trajectories) in order to create informative models that permit finding the intention or the purpose of the observed motions [4], [5]. For example, when a robot observes a human performing the activity of washing a dish, instead of learning activity models from specific *low-level* parameters such as velocity or positions, the robot extracts the meaning of washing using abstract representations [6]. This allows the learned semantic description of the activity of washing to be used in a different scenario since the obtained model allows for this type of generalization.

One of the main advantages of generating semantic-based models for either activity recognition or action execution is to create formal descriptions to extract the syntax and semantics of human activities. However, one of the main challenges to assess the obtained semantic models is to find multi-purpose datasets. The main contribution of this extended abstract is to present a new dataset, named Household Activities from Virtual Environments (HAVE). The HAVE dataset contains three different household scenarios: dining room, kitchen and living room, which are used to evaluate different tasks, e.g. setting a table, washing dishes, and cleaning a room. Fig. 1 presents some examples of the virtual scenarios presented in this dataset.

II. Related Work

One key element to advance the research on the understanding of human movements in robotics is the used of benchmarking datasets. Huang et al [7] presented a very comprehensive analysis of the different datasets related to the topics of object manipulation and grasping. There have

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

[2] Technical University of Delft, Cognitive Robotics, The Netherlands t.bates@tudelft.nl

been many attempts to propose a common benchmarking dataset for the recognition of human activities, however, the main problem lies in the availability of the datasets since there are only few that are publicly available for testing semantic-based methods for robotic applications. For example, De la Torre. et al [8] presented the Carnegie Mellon University cooking dataset (CMU-MMAC[1]), which contains multimodal information of 5 external cameras, 1 wearable camera, 5 microphones, a Vicon motion capture system with 12 infrared cameras, Internal Measurement Units (IMUs), and wearable devices (eWatch). This dataset captures the movements of 43 subjects performing five different cooking tasks such as preparing brownies, pizzas, sandwiches, salads, and scrambled eggs. This is a very complex dataset which allows various analysis of the subject's movements captured by multiple sensors. Another dataset that includes multiple cameras is the TUM cooking dataset[2] [9], which contains three RGB cameras located in different positions and one wearable gaze camera with attached markers mounted on the head of the subjects. With the TUM dataset, it is possible to analyze the performed task from different angles and different perspectives. This dataset recorded 9 subjects for two different cooking tasks such as making a pancake and preparing a sandwich. The interesting component of this dataset is the first person view perspective, which is the view that robots have when executing the desired task. With this dataset, it is possible to analyze the focus of attention of the participants and transfer this to robots. There have been some efforts to also capture the movements of the human body, for example, the KIT Whole-Body Human Motion[3] Database [10]. The KIT dataset uses a motion capture system which is very useful for human motion analysis, imitation learning, action recognition, and motion generation in robotics. This large-scale dataset contains whole-body human motions from 38 subjects performing a wide variety of manipulation tasks such as drinking, shaking, pouring, throwing objects, preparing dough, dancing, etc. This is one of the largest datasets available in robotics which, besides the manipulative actions, also contains locomotion and gestures. The Cornell Activity Dataset[4] [11] contains recordings from two datasets, the CAD-60, and the CAD-120. Both have RGB-D video sequences of 8 humans performing different tasks in different environments such as office, kitchen, bedroom, and living room. The human movements were recorded using a Microsoft Kinect sensor for the following tasks: making cereal, brushing teeth, taking medicine, stacking objects, microwaving food, etc. The Cornell dataset also provides information about the tracked human skeletons and the annotations for tasks, activities, and affordances. Another dataset that also includes semantic annotations of the manipulation activities is the MANIAC dataset[5] [12]. The MANIAC dataset recorded 8 different manipulation activities such as pushing, stirring, cutting, chopping, etc. All manipulations were recorded using the Microsoft Kinect sensor to observe 5 different subjects performing 8 different tasks and each task was performed 15 times.

[1] http://kitchen.cs.cmu.edu/
[2] http://web.ics.ei.tum.de/∼karinne/Dataset/dataSet.html
[3] https://motion-database.humanoids.kit.edu/
[4] http://pr.cs.cornell.edu/humanactivities/
[5] https://alexandria.physik3.uni-goettingen.de/cns-group/datasets/maniac/

III. Household Activities from Virtual Environments (HAVE) Dataset

Virtual Reality (VR) is a viable way to collect realistic information about human activities in a structured manner. Fig. 1 shows the three different virtual scenarios that were developed for the acquisition of human demonstrations.

The virtual reality system runs under Windows and was built with the Unity Game Engine. Users can interact with the environment through the Vive HMD (Head Mounted Display) and its respective controllers. The movements of the users in the VR environment are stored in a log file which contains information about the 3D position, orientation (quaternion), and velocities of the end-effectors (hands of the human while using the controllers), as well as the 3D positions and the names of all the objects in the virtual scene. The proposed VR system is compatible with the Robot Operating System (ROS) which runs under Linux. The output of the VR demonstrations communicates with ROS via a WebSocket connection through the ROSBridge Server[6]. Therefore, the VR system sends information about the world's physical state such as hand movements and object information to the ROS engine for further semantic segmentation and interpretation of the human demonstrations (see Sec. IV).

The HAVE dataset contains human demonstrations of everyday tasks such as setting a table, washing dishes, and cleaning a room. This data was collected from random participants attending the Automatica[7] trade fair that took place in Munich in June 2018. Automatica is the leading exhibition for smart automation and robotics in Germany. The dataset contains performances from different types of participants ranging from teenagers ($>$ 16 years old) to adults ($<$ 70 years old).

One of the main goals of the collection of this new dataset was to obtain as many variations as possible for executions of the same task. In total, we captured 240 recordings for the three scenarios. As expected, not all the collected data is usable for the extraction of semantic representations. Fig. 2 shows some examples of four different participants. First, we see that the participant does the requested task of washing the dishes (Fig. 2a). However, the second participant (Fig. 2b) decided not to perform the requested task and instead placed the plate inside the microwave for a cooking task. The third participant has successfully set the table from the dining room (Fig. 2c). We can observe that the fourth participant is not performing the requested task, but he/she is throwing the dishes through the window (Fig. 2d). This means that not all the recorded data is usable. In consequence, we have manually labeled good and bad performances. Table I

[6] https://github.com/RobotWebTools/rosbridge_suite
[7] https://automatica-munich.com

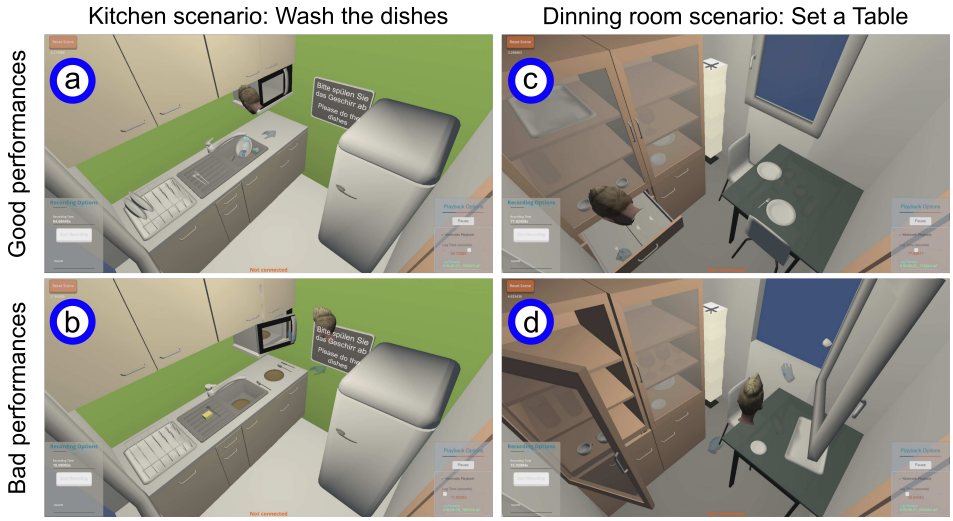

Fig. 2. Examples of the type of data obtained from the different participants. Sometimes, the participants did not perform the requested task. In such cases, the data is considered a bad example and not further analyzed.

summarizes the types of performances that we have obtained from the HAVE dataset. Overall, more than 85% of the collected data can be used for further analysis.

TABLE I
SUMMARY OF USABLE DATA FROM THE HAVE DATASET

Analyzed task	Good	Bad	Total
Wash the dishes	82	14	96
Set a table	79	4	83
Clean a room	44	17	61
Total number of recordings			240

IV. ASSESSING THE HAVE DATASET TO EXTRACT SEMANTIC REPRESENTATIONS

A human can set a table regardless of its size, the kind of dinnerware, and the shape of the glasses. We, humans, manage to do that since we can adapt to different environments [13]. Therefore, the goal of using the data collected from the HAVE dataset is to transfer the human demonstrations to service robots so that they can learn from possible variations within an environment. The collected dataset gives the possibility of analyzing different styles of performing the same task. For example, from the 79 good performances of setting a table, we observe different results. For instance, some people decided to set the table for one person, and others for two persons. Furthermore, the table was set-up in various configurations since some participants decided to use only plates and glasses to finish the task as soon as possible, whereas other participants executed the task in more detail using most of the available objects in the scene. Also, the spatio-temporal relationships between the objects vary. For example, some participants placed the knife to the right side of the plates and other participants to the left side. Therefore, the dataset captures different variations as expected, making this a unique component of the HAVE dataset.

In order to evaluate the proposed dataset, we have analyzed the data of some participants during the task of washing dishes (see Fig. 3). In our previous work [2], we presented a hierarchical learning method to continuously segment the human motions and simultaneously classifying known actions while learning new ones on demand. Our proposed learning method is a semantic-based approach that extracts the meaning of the demonstrations by means of symbolic and semantic representations. The *lowest level* of our hierarchical method finds the relevant information from the demonstrations from multiple sensors [9]. We have previously demonstrated that the relevant information is based on the velocity of the user's hand and the relative distance between the objects in the scene and the user's hand. Fig. 3b) shows that for the VR demonstrations, the system deals with multiple objects. In this case, the segmentation method discards objects that are far away from the user's hand, and only considers the objects that are closer to the hands for recognition of the demonstrated activities. This obtained information represents the input to the *highest level*, which infers the demonstrated activities using the automatically extracted semantic representations. The semantic representations are obtained using a decision tree based on the C4.5 algorithm [14]. The obtained rules are enhanced with a knowledge representation module. In this case, we use KnowRob [15] as a baseline ontology. Then, the proposed semantic-based method is able to track, segment, and recognize the movements from the user *online* [6], [2]. One important capability of this semantic-based approach is the possibility of learning new activities on-demand. Fig. 3d) shows that there are at least nine new activities that were detected during the washing dishes task. One of the limitations of this method is that it is not able to give a proper label to the newly recognized activity. In this case, all the newly detected activities are labeled as *GranularActivity_XXXXX*, where XXXXX indicates a randomly generated unique ID. As future work, we will investigate

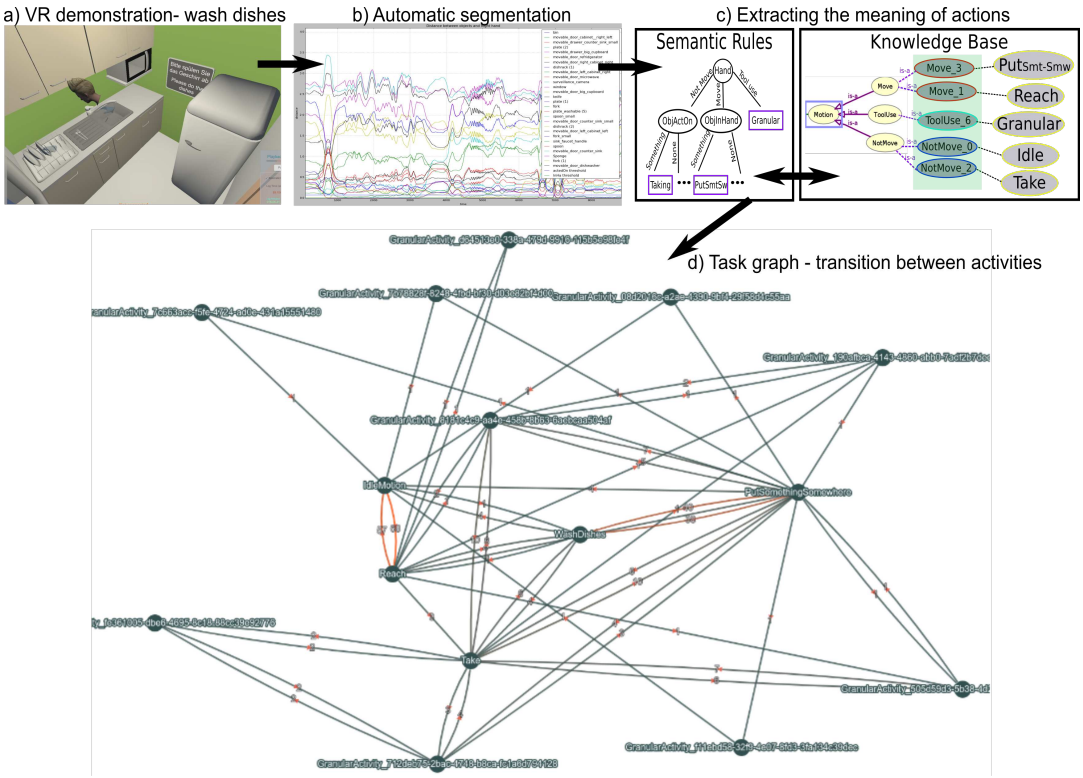

Fig. 3. Pipeline to extract semantic representations for one subject performing the washing dishes task. a) shows the end of the washing dishes demonstration in VR, b) exemplifies the challenge of having the information from all the objects in the scene for performing the automatic segmentation, c) shows that the extraction of the meaning of the activities uses semantic rules and a knowledge-base module, and finally, d) shows the generated task graph obtained at the end of the demonstration.

if the newly generated activities are correct before labeling them correctly.

The main goals of semantic-based approaches are to enable standard robots to be flexible, modular, and adaptive to different environments. In order to test all these different capabilities, rich datasets are needed. For example, to test the flexibility of the state-of-the-art learning approaches, it is needed to analyze the robustness of the presented method to different variations. As previously mentioned, different people perform the same task in several ways, these different demonstrations styles are captured in the HAVE dataset. Therefore, the flexibility and adaptability of the learning method can be assessed. To demonstrate this, we have tested our previously proposed semantic-based method [6] with one random participant from the HAVE dataset for the task of washing dishes. This means that the learning system has never seen this demonstration (this demonstration is not part of the training sample). The semantic-based method successfully identifies known activities while learning newly demonstrated ones. Furthermore, the learning method constructs a graph, which we called a task graph (see Fig. 3d)), of all observed activities and their relationships through continuous observations. This is later used for the robot planning process [6]. The robot understands and recognizes the activities during the demonstration in the virtual environment system and utilizes this information to execute the learned task. Thus demonstrating the major potential that the presented HAVE dataset has for testing state-of-the-art leaning methods.

V. Conclusions

One of the challenging problems in robotics is to teach robots new activities and learning methods need to cope with a new and large variety of situations. These methods should adapt to different scenarios to allow the transference of the learned models across different situations. In order to achieve that, novel and rich datasets are needed to validate the robustness and flexibility of the proposed learning methods. In this paper, we proposed a multi-purpose dataset that collected demonstrations from random participants in virtual scenarios. This dataset contains 240 recordings obtained from three different household scenarios, such as: *setting a table*, *washing dishes*, and *cleaning a room*. From this dataset, it is possible to extract the meaning of the demonstrations, which allows for identifying new activities, thus demonstrating the potential of the proposed dataset to assess the scalability and domain transfer of learning approaches.

Acknowledgments

The research reported in this paper has been (partially) supported by the German Research Foundation DFG, as part of Collaborative Research Center (Sonderforschungsbereich) 1320 EASE - Everyday Activity Science and Engineering, University of Bremen (http://www.ease-crc.org/).

REFERENCES

[1] S. Park and J. Aggarwal, "Semantic-level Understanding of Human Actions and Interactions using Event Hierarchy," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshop.*, June 2004, pp. 12–12.

[2] K. Ramirez-Amaro, M. Beetz, and G. Cheng, "Transferring skills to humanoid robots by extracting semantic representations from observations of human activities." *Artificial Intelligence*, vol. 247, pp. 95–118, 2017. [Online]. Available: https://doi.org/10.1016/j.artint.2015.08.009

[3] T. R. Savarimuthu, A. G. Buch, C. Schlette, N. Wantia, J. Robmann, D. M. Martínez, G. Alenyà, C. Torras, A. Ude, B. Nemec, A. Kramberger, F. Wörgötter, E. E. Aksoy, J. Papon, S. Haller, J. H. Piater, and N. Krüger, "Teaching a Robot the Semantics of Assembly Tasks." *IEEE Trans. Systems, Man, and Cybernetics: Systems*, vol. 48, no. 5, pp. 670–692, 2018. [Online]. Available: https://doi.org/10.1109/TSMC.2016.2635479

[4] Y. Kuniyoshi, M. Inaba, and H. Inoue, "Learning by watching : Extracting reusable task knowledge from visual observation of human performance," *IEEE Transactions on Robotics and Automation*, vol. 10, no. 6, pp. 799–822, 1994.

[5] G. Cheng, K. Ramirez-Amaro, M. Beetz, and Y. Kuniyoshi, "Purposive learning: Robot reasoning about the meanings of human activities," *Science Robotics*, vol. 4, no. 26, 2019. [Online]. Available: http://robotics.sciencemag.org/content/4/26/eaav1530

[6] 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)*. IEEE, 2017, pp. 3510–3515. [Online]. Available: https://doi.org/10.1109/IROS.2017.8206193

[7] Y. Huang, M. Bianchi, M. Liarokapis, and Y. Sun, "Datasets on object manipulation and interaction: a survey," *Big Data*, vol. 4, no. 4, Dec. 2016. [Online]. Available: https://doi.org/10.1089/big.2016.0042

[8] F. De la Torre, J. Hodgins, A. Bargteil, X. Martin, J. Macey, A. Collado, and P. Beltran, "Guide to the Carnegie Mellon University Multimodal Activity (CMU-MMAC) Database," CMU-RI-TR-08-22. Carnegie Mellon University, Tech. Rep., 2009.

[9] K. Ramirez-Amaro, H. N. Minhas, M. Zehetleitner, M. Beetz, and G. Cheng, "Added Value of Gaze-Exploiting Semantic Representation to Allow Robots Inferring Human Behaviors," *ACM Trans. Interact. Intell. Syst.*, vol. 7, no. 1, pp. 5:1–5:30, March 2017.

[10] C. Mandery, Ö. Terlemez, M. Do, N. Vahrenkamp, and T. Asfour, "The KIT whole-body human motion database," in *2015 International Conference on Advanced Robotics (ICAR)*, July 2015, pp. 329–336.

[11] H. S. Koppula, R. Gupta, and A. Saxena, "Learning Human Activities and Object Affordances from RGB-D Videos," *CoRR*, vol. abs/1210.1207, 2012.

[12] E. E. Aksoy, M. Tamosiunaite, and F. Wörgötter, "Model-free incremental learning of the semantics of manipulation actions," *Robotics and Autonomous Systems*, pp. 1–42, 2014.

[13] A. Collins and E. F. Loftus, "A spreading-activation theory of semantic processing." 1975.

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

[15] M. Beetz, D. Beßler, A. Haidu, M. Pomarlan, A. K. Bozcuoglu, and G. Bartels, "KnowRob 2.0 – A 2nd Generation Knowledge Processing Framework for Cognition-enabled Robotic Agents," in *International Conference on Robotics and Automation (ICRA)*, Brisbane, Australia, 2018.