Extension of the Action Verb Corpus for Supervised Learning

Matthias Hirschmanner[1], Stephanie Gross[2], Brigitte Krenn[2], Friedrich Neubarth[2], Martin Trapp[2], Michael Zillich[1] and Markus Vincze[1]

***Abstract*— The Action Verb Corpus (AVC) is a multimodal dataset of simple actions for robot learning. The extension introduced here is especially geared to supervised learning of actions from human motion data. Recorded are RGB-D videos of the test scene, grayscale videos from the user's perspective, human hand trajectories, object poses and speech utterances. The three actions TAKE, PUT and PUSH are annotated with labels for the actions in different granularity.**

I. INTRODUCTION

Future social robots will have to acquire new tasks and behaviors on the go through interaction with users. They need to understand scenes, natural language instructions and user motions. In order to learn new actions via imitation or verbal instructions, empirical human data is needed. We introduced the Action Verb Corpus (AVC) as a multimodal dataset with simple object manipulation actions inspired by early parent-infant communication [1]. The extension presented in this paper is focused on supervised learning for action recognition from human motion data.

Existing datasets for action recognition that provide skeleton tracking often use the Microsoft Kinect camera such as the NTU RGB+D dataset [2] or the Montalbano dataset [3]. The Kinect tracks the whole-body skeleton but lacks individual finger tracking. For the dataset provided by Marin, Dominio and Zanuttigh [4], the Kinect as well as the Leap Motion sensor were used to capture the joint positions of fingers for American Sign Language gestures.

The extension of the AVC is geared towards robotic learning of interaction with objects. The joint positions of the fingers and the object poses are tracked. The recorded manipulations of objects located on a table are annotated in two degrees of granularity. Coarse labels reflect how the users refer to the action (e.g., TAKE, PUT, PUSH). Fine labels split an action into more granular motion primitives (e.g., REACH, GRAB, MOVE OBJECT).

II. DATASET

The AVC is a multimodal dataset of simple actions for robot learning from demonstration. It was recorded from inexperienced users performing the simple actions TAKE, PUT and PUSH with different objects according to visual instructions. They were verbalizing what they were doing in German. For example, the user moves the bottle to the left side of the box and says, "Ich nehme die Flasche und stelle sie neben die Schachtel" ("I take the bottle and put it next to the box").

For the extension of the Action Verb Corpus, users experienced with the system performed the same three basic actions arbitrarily. These actions were annotated afterwards to be used for supervised learning for action recognition. This approach was chosen to obtain recordings with good tracking performance for training a machine learning model. We will use the dataset for action classification of simple actions from human motion data in order to provide the basis for robotic learning from demonstration.

A. Setup

In the basic setup, a box, a bottle and a can are positioned on a table. The user wears the Oculus Rift DK2 virtual reality headset with the Leap Motion sensor mounted on top of it. A Microsoft Kinect camera is directed at the table for object tracking. During data collection, the user moves the object on the table and describes the actions he/she is performing. The speech utterances are recorded. The setup can be seen in Fig. 1.

The Leap Motion is a stereo infrared camera constructed particularly for hand tracking. The provided software fits a hand model to the pair of captured images to retrieve the joint positions. It returns the joint position of the human hand down to the singular finger segments with sub-millimeter precision [5].

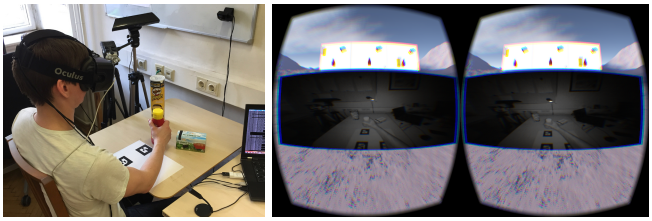

Fig. 1. The data collection setup with a user performing actions (left). Screenshot of the image shown in the Oculus Rift with the camera feed in the middle and the instructions on top (right).

The Oculus Rift headset provides the user's head pose. On the display of the headset, the user sees the scene in front of her/him as captured by the Leap Motion infrared cameras. This forces the user to direct the Leap Motion at the action she/he is performing. Therefore, the head pose can be used as an indication of gaze direction. It also ensures best possible hand tracking performance. The instructions the user has to

[1]Matthias Hirschmanner, Michael Zillich and Markus Vincze are with the Faculty of Electrical Engineering, Automation and Control Institute, Vision for Robotics, TU Wien, 1040 Wien, Austria {hirschmanner, zillich, vincze}@acin.tuwien.ac.at

[2] Stephanie Gross, Brigitte Krenn, Friedrich Neubarth and Martin Trapp are with the Austrian Research Institute for Artificial Intelligence (OFAI), 1010 Wien, Austria {stephanie.gross, brigitte.krenn, friedrich.neubarth, martin.trapp}@ofai.at

perform are displayed in the virtual reality headset above the camera images (Fig. 1).
Object tracking is performed on the monoscopic RGB images of the Kinect camera using an object tracker provided by the V4R library[1]. Models of the objects for tracking are created beforehand as described in [6]. Additionally, two binary features are saved: object is in contact with the table and object is in contact with a hand. The former is set automatically depending on the object's position, the latter is annotated manually. If the object is not in contact with a hand, averaging over consecutive object poses is performed weighted with the confidence of the object tracker because we assume the object does not move. This way, the jittering of the raw object-tracker data is reduced and occlusions do not impair tracking performance if the object was successfully tracked before.
The poses of the tracked entities (head, hands and objects) are transformed to a common coordinate frame and manually time-aligned.

B. Collected Data

The original Action Verb Corpus consists of 140 instances of TAKE/PUT actions and 110 instances of PUSH actions performed by 12 users following visual instructions. The focus is on word-object and word-action mapping.
The extension of the Action Verb Corpus consists of 210 instances of TAKE/PUT actions and 100 instances of PUSH actions performed by 2 experienced users without any instructions. The focus is on generating motion tracking data. A visualization of the tracked human arm and object poses is shown in Fig. 2. An issue in the original AVC is that the tracking information of the user's arm is lost sometimes while interacting with objects. An experienced user is able to operate the system in a way to get better tracking results and therefore more consistent data for a learning algorithm. The extension of the AVC is complementary to the original AVC.
The tracked data is annotated with action labels. Two types of annotations are created. The coarse annotation is how the user refers to the action. The classes of the coarse annotation are TAKE, PUT and PUSH. The fine annotation splits the actions into more granular motion primitives – REACH, GRAB, MOVE OBJECT and PLACE. The idea is that these primitives are more useful for the generation of robot actions while the coarse annotations reflect more complex motion concepts. For example, the robot might imitate human movement for reaching for an object. For grasping, it might switch to a different motion planner because the movement has to be adapted to the exact object pose. The coarse annotations are important for our overall goal of learning concepts of actions and link them with uttered verbs in order to acquire multimodal representations. This approach of labels with different granularity is similar to Koppula, Gupta and Saxena [7] who divide high level activities in sub-activities.

[1]https://www.acin.tuwien.ac.at/vision-for-robotics/software-tools/v4r-library/

The recordings of the extension of the Action Verb Corpus are represented by:

- 3D joint positions of the human arms, hands and fingers
- Head pose of the user
- Object poses with its corresponding confidence
- Binary features if the object touches a hand or the table
- Action annotations (coarse and fine)
- An animation of the tracked hands and objects
- RGB-D video of the scene
- Grayscale video from the user's perspective
- Recorded speech utterances

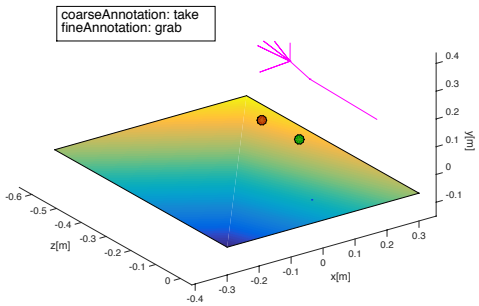

Fig. 2. Animation of the tracked data. A simplified version of the tracked arm is shown in magenta, the two objects are represented by the colored circles, the plane represents the table and the current action annotation is shown on top.

III. CONCLUSION AND FUTURE WORK

The Action Verb Corpus with its extension will be made available to the scientific community alongside this publication[2]. At the point of writing the dataset consists of 210 annotated TAKE/PUT and 100 PUSH actions. The data collection is still ongoing and will be further extended. We will use the dataset for action recognition of simple actions in order to provide the basis for robotic learning from demonstration. We want to extend the corpus with more complex actions. Additionally, we are working on alternative possibilities for human motion tracking that are less intrusive than our current setup. In a future step, a system will be implemented on a humanoid robot that will be able to detect different classes of actions and associate them with the user's utterance. Eventually, the robot should generate these actions and verbalize the imitated movements.

ACKNOWLEDGMENT

This research is supported by the Vienna Science and Technology Fund (WWTF), project RALLI – Robotic Action- Language Learning through Interaction (ICT15-045) and the CHIST-ERA project ATLANTIS (2287-N35). The Austrian Research Institute for Artificial Intelligence is supported by the Austrian Federal Ministry for Science, Research and Economy.

[2]http://ralli.ofai.at/datasets.html

REFERENCES

[1] S. Gross, M. Hirschmanner, B. Krenn, F. Neubarth, and M. Zillich, "Action verb corpus," in *Proc. 2018 Language Recognition and Evaluation Conference*, Miyazaki, Japan, May 2018.

[2] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3D human activity analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1010–1019.

[3] S. Escalera, X. Baró, J. Gonzalez, M. A. Bautista, M. Madadi, M. Reyes, V. Ponce-López, H. J. Escalante, J. Shotton, and I. Guyon, "Chalearn looking at people challenge 2014: Dataset and results," in *Workshop at the European Conference on Computer Vision*. Springer, 2014, pp. 459–473.

[4] G. Marin, F. Dominio, and P. Zanuttigh, "Hand gesture recognition with Leap Motion and Kinect devices," in *Image Processing (ICIP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1565–1569.

[5] F. Weichert, D. Bachmann, B. Rudak, and D. Fisseler, "Analysis of the accuracy and robustness of the Leap Motion controller," *Sensors*, vol. 13, no. 5, pp. 6380–6393, 2013.

[6] J. Prankl, A. Aldoma, A. Svejda, and M. Vincze, "RGB-D object modelling for object recognition and tracking," in *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*. IEEE, 2015, pp. 96–103.

[7] H. S. Koppula, R. Gupta, and A. Saxena, "Learning human activities and object affordances from RGB-D videos," *The International Journal of Robotics Research*, vol. 32, no. 8, pp. 951–970, 2013.