homer@UniKoblenz: Winning Team of the RoboCup@Home Open Platform League 2018

Raphael Memmesheimer(✉), Ivanna Mykhalchyshyna, Viktor Seib, Tobias Evers, and Dietrich Paulus

Active Vision Group, Institute for Computational Visualistics, University of Koblenz-Landau, 56070 Koblenz, Germany
{raphael,ivannamyckhal,vseib,tevers,paulus}@uni-koblenz.de
http://homer.uni-koblenz.de
http://agas.uni-koblenz.de

Abstract. We won this year's RoboCup@Home track in the Open Platform League in Montreal (Canada). The approaches as used for the competition are briefly described in this paper. The robotic hardware of our custom built robot Lisa and the PAL Robotics TIAGo, both running the same methods, are presented. New approaches for object recognition, especially the preprocessed segment augmentation, effort based gripping, gesture recognition and approaches for visual imitation learning based on continuous spatial observations between a demonstrator and the interacting objects are presented. Further, we present the current state of research of our Imitation Learning approaches, where we propose a hybrid benchmark and methods for bootstrapping actions. Furthermore, our research on point cloud based object recognition is presented.

Keywords: RoboCup@Home · Imitation learning · Gesture recognition · Object recognition · Object manipulation · RoboCup · Open platform league · Domestic service robotics · homer@UniKoblenz

1 Introduction

In this year's RoboCup we successfully participated in the RoboCup@Home Open Platform League, where we achieved the first place. After the RoboCup World Cup in Nagoya (Japan (2017)) and Hefei (China (2015)), this is the third time that we won this title. The team consisted of one supervisor and five students. Additionally, two more students were supporting the preparation.

Besides the RoboCup competitions, we also attend the European Robotics League and the World Robot Summit. For this year's participation we focused on imitation learning by observation of humans. We demonstrated this twice. Once at the RoboCup GermanOpen and once in RoboCup@Home Open Platform league. This year we also improved our team's infrastructure by a continuous

D. Holz et al. (Eds.): RoboCup 2018, LNAI 11374, pp. 512–523, 2019.
https://doi.org/10.1007/978-3-030-27544-0_42

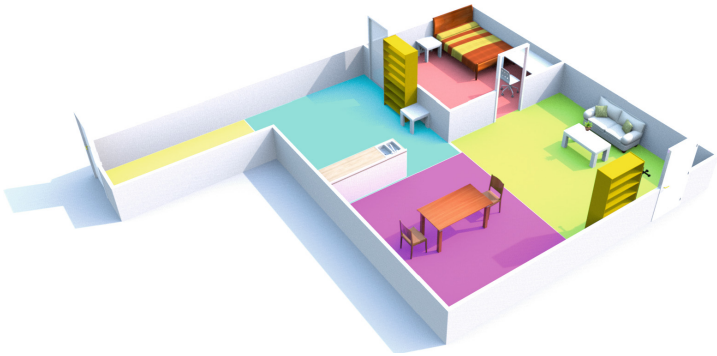


Fig. 1. The arena setup from Montreal. Different colors are corresponding to different rooms. Corridor is yellow, kitchen is cyan, dining room is pink, bedroom is red, living room is green (Color figure online)

software integration and built our packages for a variety of processor architectures.

Section 2 gives a short overview of the RoboCup@Home competition. In Sect. 3 we present the robots that we used for this year's participation. A special focus is put on the hardware changes that the robots have undergone. Section 4 describes the architecture and software approaches we are using. An overview of the current research topics is given in Sect. 5. Finally, Sect. 6 summarizes and concludes this paper.

2 RoboCup@Home

Domestic tasks are benchmarked in RoboCup@Home. The competition is divided into two stages. An *Open Challenge* and the *Finals* allow to present the current research focus in a practical application scenario. In Stage 1 focuses is on a variety of individual functionalities. The *Speech and Person Recognition* test, benchmarks the speech recognition, sound source localization, person recognition and gender estimation. In *Help Me Carry* robots are supposed to follow a person outside of the apartment to a car. The person hands over a shopping bag in a natural way. On the way back obstacles are put into the path of the robot that should be avoided. In *Storing Groceries* the robot has to pick groceries from a table and sort them into a shelf where items of the same category are already stored. In the *General Purpose* (GPSR) task all possible capabilities are tested. The robot receives a speech command consisting of several sub-commands and has to execute them. Stage 2 consists of the *Dishwasher Test*, *Restaurant*, *EE-GPSR* (Extended Endurance General Purpose Service Robot) and the *Open Challenge*. The *Dishwasher Test* focuses on precise manipulation. Robots have to open a dishwasher and store cutlery and plates safely. Further, a dish washing tab should be placed into the dishwasher. The *Restaurant* task takes place outside of the arena in a previously unknown restaurant. The robots are placed

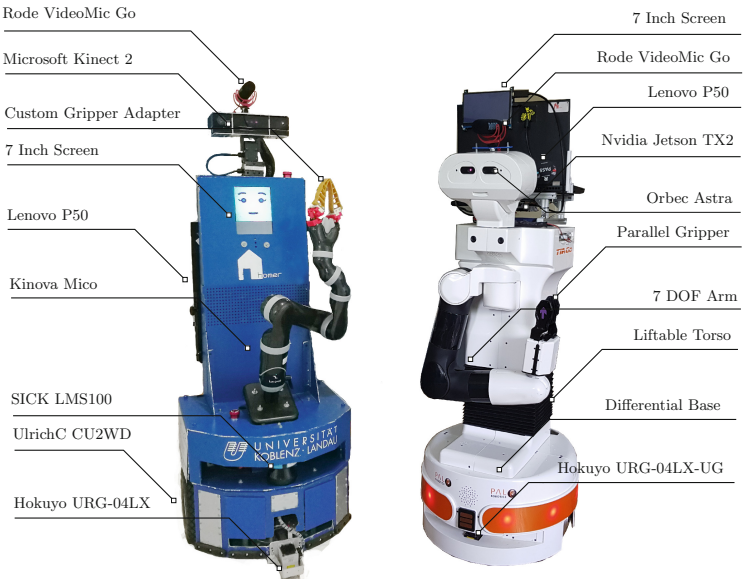


Fig. 2. The robots Lisa (left) and Marge (right). Lisa is our main robot inspired by Marge as a successor. Both robots run the same software with minor exceptions like the model descriptions and hardware interfaces.

at an initial location and search for a waving or shouting person. A map in a dynamic environment needs to be created online. The ordering persons should be approached and asked for an order. This is particularly hard as the chosen scenarios are usually quite noisy. The *EE-GPSR* task is an enhanced-endurance version of the GPSR task where multiple robots are operating at the same time. The arena setup from this year's RoboCup in Montreal is shown in Fig. 1.

3 Hardware

We use a custom built robot called Lisa and a PAL Robotics TIAGo (Marge). Lisa is built upon a CU-2WD-Center robotics platform. The PAL Robotics TIAGo robot is able to move its torso up and down and has a wider working range. Currently, we are using a workstation notebook equipped with an Intel Core i7-6700HQ CPU @ 2.60 GHz × 8, 16 GB RAM with Ubuntu Linux 16.04 and ROS Kinetic. Each robot is equipped with a laser range finder (LRF) for navigation and mapping. The most important sensors of Lisa are set up on top of a pan-tilt unit. Thus, they can be rotated to search the environment or take a better view of a specific position of interest. Apart from a RGB-D camera (Microsoft Kinect 2) a directional microphone (Rode VideoMic Pro) is mounted on the pan-tilt unit. A 6-DOF robotic arm (Kinova Mico) is used for mobile manipulation. The end effector is a custom setup and consists of 4 Festo Finray-fingers. Finally, an Odroid C2 inside the casing of Lisa handles the robot face

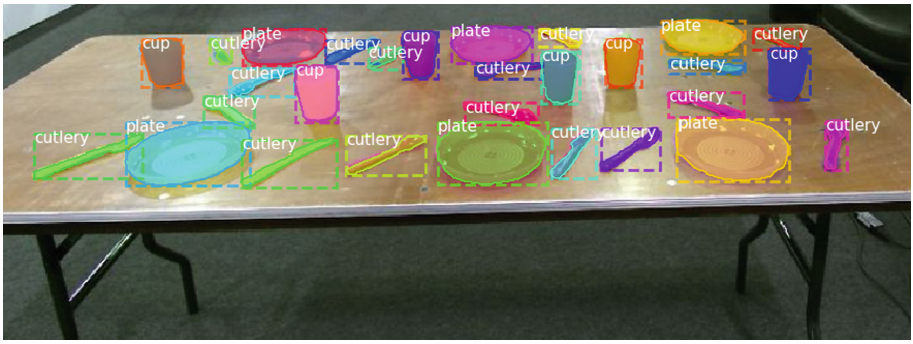


Fig. 3. The Mask-RCNN resulting image of cutlery objects as used for the Dishwasher challenge.

and speech synthesis. A Raspberry Pi 3 in combination with a Matrix Creator board was used for the integration of a sound source localization system.

TIAGo has a mobile differential base with a Hokuyo URG-04LX-UG01 laser range finder. We compute the odometry from wheel encoders. In combination with the LRF we can create a map of the environment and localize the robot. The robot has a torso which is lift-able by 40 cm. In case of the toilet cleaning task during the World Robot Summit competition this was a benefit for the use of the sponge-end effector. We could reach the toilet seat with a top-down end-effector pose, but also the ground of the toilet. The 7-DOF arm was used for cleaning the toilet seat, picking up the paper pieces and cleaning the floor of the toilet. The head has a 2-DOF and holds an Orbbec ASTRA RGB-D camera which was used for the segmentation of the toilet seat and the detection of the trash on the floor. Further, we used a Lenovo P50 workstation notebook equipped with Intel i7-6700HQ CPU, 16 GB memory and a 2 GB Quadro M1000 GPU and mounted it on the back of TIAGo. A NVIDIA Jetson TX2 was used to compensate for the low graphical memory available on TIAGO's notebook in order to run multiple models in parallel. The robot setup is depicted in Fig. 2.

4 Approaches

This section briefly introduces our approaches. The applied software architecture has been described previously [1–3].

Object Recognition. This year we used Mask-RCNN [4] in combination with a custom augmentation approach as a segmentation method for images. The segmented images with the labels are augmented in image space among different backgrounds of the arena. In the background images we ensured that no relevant objects are visible. The use of background images decreases also the false positive detections. This segmentation method is beneficial for more precise manipulation

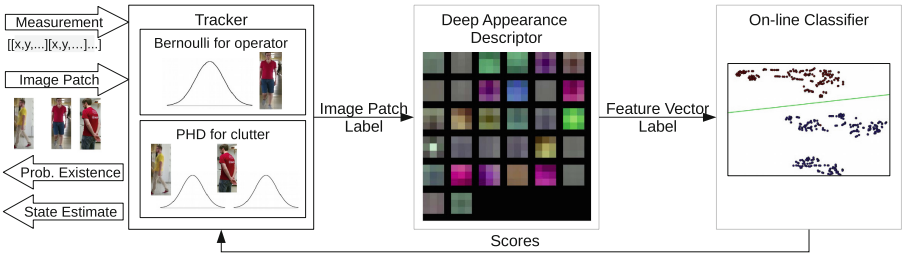


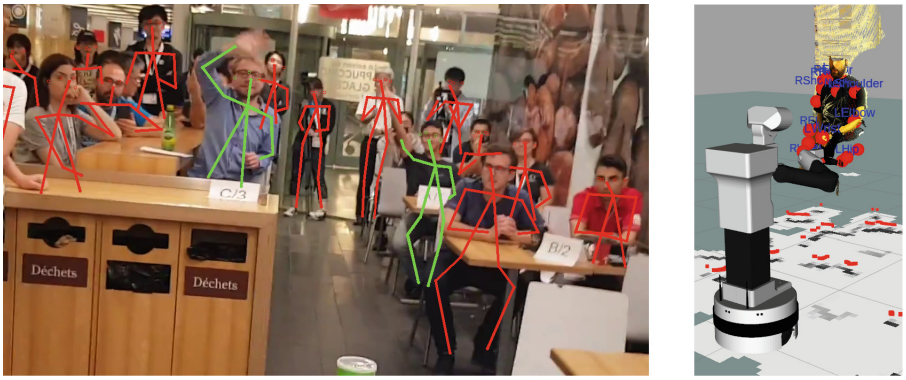
Fig. 4. Tracking Overview. We employ a RFS Bernoulli single target tracker in combination with a deep appearance descriptor to re-identify and online classify the appearance of the tracked identity. Measurements, consisting of positional information and an additional image patch serve as input. The Bernoulli tracker estimates the existence probability and the likelihood of the measurement being the operator. Positive against negative appearances are continuously trained. The online classifier returns scores of the patch being the operator.

tasks i.e. for cutlery objects which are not segmentable in the depth images as the height differences are below the separable threshold. As the segmentation method is computationally expensive we calculate the mask proposals on single images only. A faster approach that yields object masks with high frequency is desirable as future work to allow closed loop manipulation. An exemplary segmentation image is shown in Fig. 3. In total 344 images containing containing multiple objects where labeled.

Speech Recognition. For speech recognition we use a grammar based solution supported by an academic license for the VoCon speech recognition software by Nuance[1]. We combine continuous listening with begin and end-of-speech with the integrated detection to get good results even for complex commands. Recognition results below a certain threshold are rejected. The grammar generation is supported by the content of a semantic knowledge base that is also used for our general purpose architecture.

Operator Following. We developed an integrated system to detect and track a single operator that can switch *off* and *on* when it leaves and (re-)enters the scene [5]. Our method is based on a set-valued Bayes-optimal state estimator that integrates RGB-D detections and image-based classification to improve tracking results in severe clutter and under long-term occlusion. The classifier is trained in two stages. First, we train a deep convolutional neural network to obtain a feature representation for person re-identification. Then, we bootstrap an online classifier that discriminates the operator from remaining people on the output of the state-estimator (Fig. 4). The approach is applicable for following and guiding tasks.

[1] http://www.nuance.com/for-business/speech-recognition-solutions/vocon-hybrid/index.htm.



(A) (B)

Fig. 5. Gesture recognition output: (A) extracted person poses in the *Restaurant* challenge at RoboCup@Home 2018 in Montreal, Canada: detected human poses are marked red, calling persons are highlighted green; (B) projection of the body keypoints into 3D space. (Color figure online)

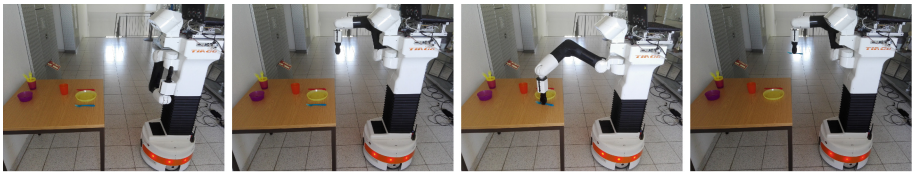
Person Detection. For person detection we integrated multiple approaches for different sensors that can be optionally fused and used to verify measurements of other sensors. A leg detector [6] is applied on the laser data. This yields high frequency, but error prone measurements. For finding persons in point clouds we follow an approach by Munaro et al. [7]. The most reliable detections are by a face detection approach [8], assuming that the persons are facing the camera. For gender estimation we then apply an approach by Levi et al. [9].

Gesture Recognition. In this section, we describe the gesture recognition approach as used during RoboCup@Home 2018 in Montreal, Canada. This method differs from our model based approach as presented in [10]. Therefore we give a more detailed description of our approach here. Gesture recognition, and in particular the waving gesture detection, is one of the features that are tested in many RoboCup challenges such as in *Restaurant*, *GPSR* and *EEGPSR*. Human pose features are extracted by Convolutional Pose Machines (CPM) [11] with pre-trained COCO-model from a RGB-image. The extracted features result in the set F with 18 possible body parts represented in the pixel space. We denote the joint keypoint as $\boldsymbol{f}_i \in F$ where:

$$\boldsymbol{f}_i = \begin{bmatrix} x_i \\ y_i \end{bmatrix}, \tag{1}$$

and $i \in \{0, ..., 17\}$. Moreover, we define a vector connecting i_{th} and j_{th} body parts as follow:

$$\boldsymbol{v}_{i,j} = \begin{bmatrix} x_i - y_j \\ y_i - y_j \end{bmatrix}, \tag{2}$$



(A) (B) (C) (E)

Fig. 6. Illustration of our effort gripping approach: (A) is the starting position. In (B) the robot prepared to grasp. (C) the arm is moved over the object and downwards while observing the torque and then stopped to grasp. In (D) the arm is lifted again and the success or failure of the grasping is verified.

with $i, j \in \{0, ..., 17\}$ and $j \neq i$. Furthermore, the angle between the vectors of two body parts is calculated exploiting the following formula:

$$a(\boldsymbol{v}_{i,j}, \boldsymbol{v}_{l,j}) = \arccos(\frac{\boldsymbol{v}_{i,j} \bullet \boldsymbol{v}_{l,j}}{||\boldsymbol{v}_{i,j}|| \cdot ||\boldsymbol{v}_{l,j}||}), \tag{3}$$

where $i, j, l \in \{0, ..., 17\}$ and $j \neq i \neq l$. Finally, the given pose is classified as the waving gesture considering angles between particular body parts defined by COCO pose format: *hand-elbow* with connecting vectors $\boldsymbol{v}_{4,3}$ for the right arm and $\boldsymbol{v}_{7,6}$ for the left arm and *elbow-shoulder* with the connecting vector $\boldsymbol{v}_{5,6}$. The angles between *hand-elbow* and *elbow-nose* with connecting vectors $\boldsymbol{v}_{0,3}$ for the right arm and $\boldsymbol{v}_{0,6}$ for the left arm are examined by utilizing Eq. 3 in the following function:

$$call() = \begin{cases} 1, & \text{if } 0 < a(\boldsymbol{v}_{4,3}, \boldsymbol{v}_{0,3}) \leq \theta \wedge 0 < a(\boldsymbol{v}_{4,3}, \boldsymbol{v}_{2,3}) \leq \theta \\ 1, & \text{if } 0 < a(\boldsymbol{v}_{7,6}, \boldsymbol{v}_{0,6}) \leq \theta \wedge 0 < a(\boldsymbol{v}_{7,6}, \boldsymbol{v}_{5,6}) \leq \theta \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

Based on conducted experiments we found that $\theta = 150$ works well for the $call()$ function defined in Eq. 4. The result of the call detection is depicted in Fig. 5(A), where the waving gesture is highlighted in green. To estimate a final position of the gesticulating person we project the average over the body part position into map-coordinates.

The result of the projection is shown in Fig. 5 (B).

A video of this approach during the Restaurant task is available[2]. A model based approach [10] has been proposed later. Currently, we are also working on an extension of the gesture recognition approach to image sequences.

Effort Gripping. For gripping tiny objects that are hardly differentiable from the underlying surface in the depth image the estimation of a precise grasp pose is not possible. This is the case for i.e. cutlery or dishwasher tabs. We therefore propose an closed loop effort (in motor-current or force-torque) based

[2] Restaurant challenge video: https://www.youtube.com/watch?v=31Tmmhhqo_4.

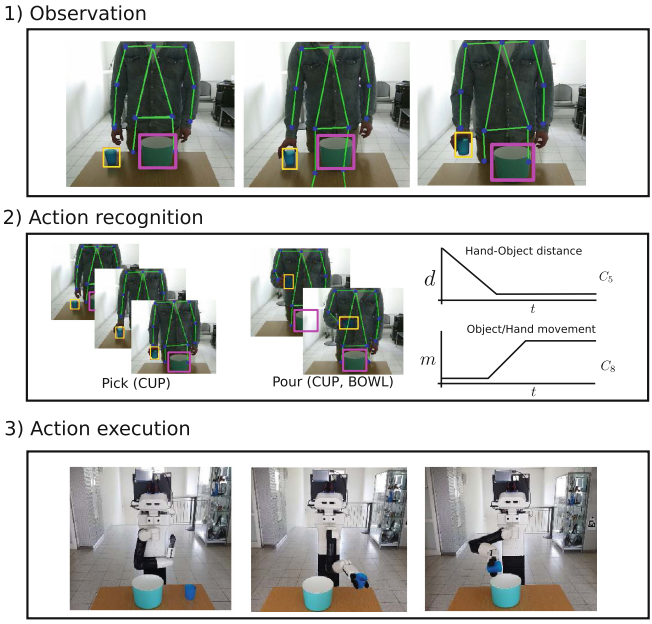


Fig. 7. Approach overview for extracting action informations from 2D image sequences in order to execute them on a mobile robot. Exemplary object detections (yellow, pink) and human pose estimates (green) are observed. Actions are recognized using a set of constraints. For replicating the observed actions we used two mobile robots equipped with an arm. (Color figure online)

grasping approach. An object pose slightly above the object position and the end-effector facing downwards is defined as an initial arm pose. Then the joint efforts are continuously observed while the end-effector is moved downwards approaching the object with Cartesian movements until the joint effort peaks. Freely spoken this approach moves the end-effector downwards to the object until the underlying surface is touched. After the object grasping we observe the joint positions of the gripper in order to verify if the object was grasped successfully. A sequential overview is given in Fig. 6. This approach has proven to be beneficial for small objects multiple times during the challenge and can also be used for safely placing objects on detected surfaces. A video of this approach integrated in the *Dishwasher* scenario is available[3].

Imitation Learning. This year in the *Open Challenge* and in the *Finals* we presented a novel approach for imitating human behavior based on visual information of a RGB camera. The human hand and objects are continuously detected. We proposed a visual approach for Imitation Learning [12]. This approach was

[3] Effort based gripping approach as used for the Dishwasher challenge: https://www.youtube.com/watch?v=luSMEtMoX7w.

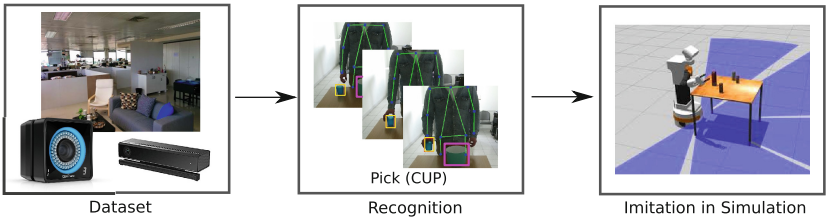


Fig. 8. Overview: this figure gives an overview of our hybrid benchmarking model. We provide a dataset recorded with a RGB-D camera and a motion capturing system. The sequences of the dataset are supposed to be interpreted by approaches for imitation learning, which then have to execute the imitation in a simulated environment grounded by the ground truth initial object positions. After the performance in simulation, results are automatically evaluated by provided scripts.

presented during the 2018 RoboCup@Home *Finals* in Montreal. Current robotic systems that lack a certain desired behavior commonly need an expert programmer to add the missing functionality. Contrary, we introduce an approach related to programming robots by visual demonstration that can be applied by common users. Provided a basic scene understanding, the robot observes a person demonstrating a task and is then able to reproduce the observed action sequence using its semantic knowledge base. We presented an approach for markerless action recognition based on Convolutional Pose Machines (CPM) [13], object observations [14] and continuous spatial relations. The actions are executable on a robot that is able to execute a set of common actions. The initial scene analysis allows semantic reasoning in case the required object is not present. Further, this allows executing the same action sequence with different objects which is a major benefit over action sequencing approaches that rely on positional data only. Even though we are demonstrating our approach on 2D observations, the formulations are also adaptable for 3D. Figure 7 gives an overview of our approach. More information is available on our project page[4].

5 Current Research

The current focus of research is on Imitation Learning. The previously described visual imitation learning approach is based on spatial relations between the demonstrator and objects and uses a mapping between recognized actions and robot actions. Additionally, we focus on research in benchmarking Imitation Learning and bootstrap actions by observation. Further, we introduce our research activities in point cloud based object recognition methods.

[4] Imitation Learning project page: https://userpages.uni-koblenz.de/~raphael/project/imitation_learning/.

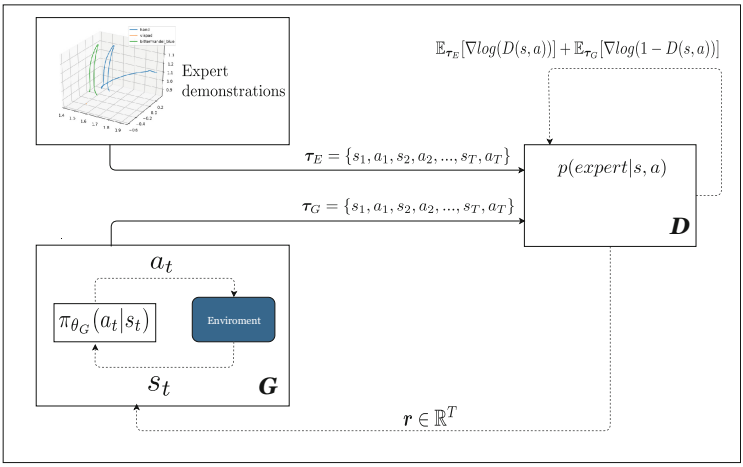


Fig. 9. Diagram of the Generative Adversarial Imitation approach for learning robotics tasks from the expert demonstrations. Two neural networks, the Discriminator network D and the Generator network G, are trained resulting in an optimal policy $\pi_\theta(a|s)$. D uses the expert trajectories $\boldsymbol{\tau}_E$ for the training to distinguish the expert trajectories from the trajectories $\boldsymbol{\tau}_G$ produced by the policy. D outputs the reward vector $\boldsymbol{r}$ which is used to optimize the policy in the inner loop.

Imitation Learning Benchmark. Currently, there is a lack of datasets for Imitation Learning through human observation. We created a benchmark called *Simitate* where we recorded a dataset using a commonly available RGB-D camera calibrated against a motion capturing system. Data was recorded in a domestic real world testbed as used for the European Robotics League. Different people performed daily activities like performing movements with their hand, picking, placing, stacking or moving objects. For the demonstrator's hand and the interacting objects ground truth poses are recorded with the motion capturing system. As ground truth positions of the objects and hand are available we can spawn them into a simulated representation of the environment where simulated robots should imitate the demonstrations. We suggest metrics for effect and trajectory level imitations. The approach is visualized in Fig. 8 and more information can be found on the project page[5].

General Adversarial Imitation Learning. Generative Adversarial Imitation Learning (GAIL) was recently proposed by Ho and Ermon [15] as an approach to teach a robot to accomplish the given task using expert demonstrations. In our work we leverage from GAIL and regard the robotic manipulation problem as a sequential decision making task where the robot follows a stochastic parametrized policy $\pi_\theta(a|s)$ that maps observed state s to a distribution over

[5] Simitate Imitation Learning project page: https://agas.uni-koblenz.de/data/simitate/.

manipulation actions a. The overview of the approach is depicted in the Fig. 9. For training the generative model we use the expert trajectories from the dataset described in Sect. 5. Furthermore, we exploit the Proximal Policy Optimization method [16] in the inner loop as it is shown in the Fig. 9 by utilizing reward vector $\boldsymbol{r} \in R^T$ given by the Discriminator in order to estimate the advantage function.

Point Cloud Recognition and Affordance Estimation. Despite the great success of deep learning approaches in the recent years, we also continue research on some classic methods. Classic methods are especially useful if only little training data or computational resources are available. We further refined the nearest-neighbor point cloud recognition presented in [17] to be computationally more efficient and achieve higher classification rates with an optimized codebook filtering method[6]. This approach will be further combined with an affordance estimation algorithm [18] in the future.

6 Summary

RoboCup@Home is a robot competition where domestic robots compete in daily tasks. For this year we used two robots, a custom build robot called Lisa and a PAL Robotics TIAGo. We briefly described our approaches and presented our research focus. Novel approaches for gesture recognition, effort based gripping and visual imitation learning where successfully presented in this year's competition. In the finals we further demonstrated that the acquired knowledge by one observing robot is reusable by other robots with a common set of functionalities.

Acknowledgement. First we want to thank the participating students Niklas Yann Wettengel, Tobias Evers, Lukas Buchhold, Thies Möhlenhof, Lukas Debald and Anatoli Eckert. A major thanks is given to PAL Robotics SL which supported us with the free loan of a TIAGo robot and further supported us in organizing the shipping to Montreal. Thanks to Nuance Communications Inc. for supporting the team with an academic license for speech recognition. Further, we want to thank NVIDIA for the grant of a graphics card that has been used for training the operator re-identification and the object segmentation.

References

1. Memmesheimer, R., Seib, V., Paulus, D.: homer@UniKoblenz: winning team of the RoboCup@Home open platform league 2017. In: Akiyama, H., Obst, O., Sammut, C., Tonidandel, F. (eds.) RoboCup 2017. LNCS (LNAI), vol. 11175, pp. 509–520. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00308-1_42
2. Memmesheimer, R., et al.: Robocup: homer@unikoblenz (germany). Fachbereich Informatik. Technical report 4/2018 (2018)

[6] Detailed description and code on https://github.com/vseib/PointCloudDonkey.

3. Seib, V., Memmesheimer, R., Paulus, D.: A ROS-based system for an autonomous service robot. In: Koubaa, A. (ed.) Robot Operating System (ROS). SCI, vol. 625, pp. 215–252. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-26054-9_9
4. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2980–2988. IEEE (2017)
5. Wojke, N., Memmesheimer, R., Paulus, D.: Joint operator detection and tracking for person following from mobile platforms. In: 2017 20th International Conference on Information Fusion (Fusion), pp. 1–8, July 2017
6. Lu, D.V., Smart, W.D.: Towards more efficient navigation for robots and humans. In: IEEE/RSJ International Conference On Intelligent Robots and Systems (IROS), pp. 1707–1713. IEEE (2013)
7. Munaro, M., Menegatti, E.: Fast RGB-D people tracking for service robots. Auton. Robots **37**(3), 227–242 (2014)
8. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process. Lett. **23**(10), 1499–1503 (2016)
9. Levi, G., Hassner, T.: Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, pp. 503–510. ACM (2015)
10. Memmesheimer, R., Mykhalchyshyna, I., Paulus, D.: Gesture recognition on human pose features of single images. In: 2018 9th International Conference on Intelligent Systems (IS), pp. 1–7. IEEE (2018)
11. Wei, S.-E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4724–4732 (2016)
12. Memmesheimer, R., Mykhalchyshyna, I., Seib, V., Theisen, N., Paulus, D.: Markerless visual robot programming by demonstration. CoRR, vol. abs/1807.11541 (2018). http://arxiv.org/abs/1807.11541
13. Wei, S., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. CoRR, vol. abs/1602.00134 (2016). http://arxiv.org/abs/1602.00134
14. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)
15. Ho, J., Ermon, S.: Generative adversarial imitation learning. In: Advances in Neural Information Processing Systems, pp. 4565–4573 (2016)
16. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms.' arXiv preprint arXiv:1707.06347 (2017)
17. Seib, V., Link, N., Paulus, D.: Pose estimation and shape retrieval with hough voting in a continuous voting space. In: Gall, J., Gehler, P., Leibe, B. (eds.) GCPR 2015. LNCS, vol. 9358, pp. 458–469. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24947-6_38
18. Seib, V., Knauf, M., Paulus, D.: Affordance origami: unfolding agent models for hierarchical affordance prediction. In: Braz, J., et al. (eds.) VISIGRAPP 2016. CCIS, vol. 693, pp. 555–574. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-64870-5_27