Imitating Task-oriented Grasps from Human Demonstrations with a Low-DoF Gripper

Timothy Patten and Markus Vincze

Automation and Control Institute
TU Wien
Vienna, Austria
Email: {patten, vincze}@acin.tuwien.ac.at

***Abstract*—Task-oriented or semantic grasping is important in robotics because it enables objects to be manipulated appropriately and used for their intended purpose. Many objects are human designed, therefore, we address the problem of learning task-oriented grasps by directly observing human behaviour. A person simply demonstrates the appropriate grasp, which is quick and convenient for any user in the real world. Our approach uses RGB images to track the object and hand pose, then employs a neural network to translate the human hand configuration to a robotic grasp with fewer degrees of freedom. Analysis shows that a variety of low-dimensional representations of the hand enable the mapping to be learned and that the model better generalises to new demonstrators handling new objects when the training data is augmented. Experiments with a mobile manipulator show that a robot successfully observes grasps and imitates the action on objects in various poses. This is accomplished immediately, without additional learning and is robust in real-world conditions.**

Keywords–Robotic grasping; task-oriented grasping; learning from demonstration; imitation learning; deep learning.

I. INTRODUCTION

Grasping is an important capability for robots operating in industry or in homes. The human world is very complex: objects have a variety of characteristics and the environment imposes unpredictable constraints. As such, learning generalisable grasping strategies that are robust in real-world conditions is an ongoing and active field of research.

Significant advances have been made by applying deep learning, which has been enabled by the introduction of large datasets that are annotated by hand [1] [2], compiled with 3D object models and analytical metrics [3] [4] or generated using grasp planners in simulation [5]–[7]. Grasping methods trained on these datasets ignore the semantics of the grasp and only measure success if an object is securely lifted or transported. Semantic or task-oriented grasping introduces the concept that objects should be grasped to enable task-related manipulation actions [8]. For example, grasping the handle and not the blade enables a knife to be used for cutting. Existing approaches exploit manually annotated examples [9] [10], constrain grasps to parts that afford the task [11]–[13] or perform self-supervised learning in simulation [14].

In this paper, we address task-oriented grasping by Learning from Demonstration (LfD) in which a robot learns to replicate a grasp demonstrated by a human, as shown in Figure 1. In contrast to previous work, we contribute a convenient human-robot interface that requires no special instrumentation [15], manual annotation [16], physical interaction with the robot [13] or an offline learning process [17]. Our system uses only the onboard camera to observe a single example of a human grasp to understand how the grasp is performed, which makes it easy to be adopted by untrained users in the real world. Furthermore, we explicitly address the problem of transferring the human grasp to a robot gripper with fewer Degrees of Freedom (DoF). We employ a neural network to learn the mapping between the joints of the human hand and the parameterisation of the robot grasp to effectively transfer the observed human grasp to a robotic parallel-jaw gripper.

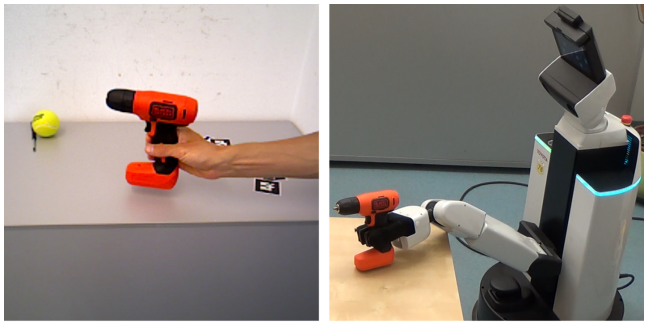


Figure 1. Human demonstrates how to semantically grasp an object by lifting a drill by the handle (left). Robot imitates the semantic grasp (right).

An analysis of the neural network shows that it has the capacity to successfully transfer human grasps to the robot platform. In an ablation study, we show that it is sufficient to learn from a subset of hand joints to yield high quality robot grasp pose predictions. Furthermore, learning from interactions with one object better transfers to other objects when the training data is augmented. In experiments with a mobile manipulator, people demonstrate semantic grasps by grasping the handles of objects, such as mugs and drills. The robot shows the ability to observe the human and instantly imitate the demonstrated grasp on the same object when it is presented in any new pose.

In summary, we make the following contributions:

- A neural network architecture to regress the grasp parameterisation for a low-DoF robotic gripper from the human hand configuration;
- evaluation of the grasp regression network for transferring between different demonstrators and objects;
- a grasp imitation learning pipeline using state-of-the-art object pose estimation and hand tracking with our regression network to transfer demonstrated human grasps to robot grasps for the observed objects; and
- experiments of real-world task-oriented grasp learning from demonstration with a mobile manipulator.

 ISBN: 978-1-61208-787-0

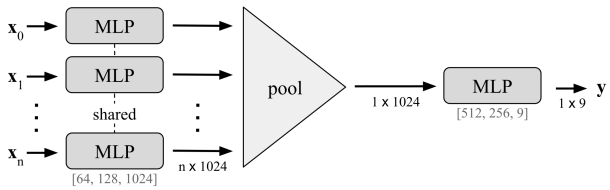
The remainder of the paper is organised as follows. Section II discusses related work. Section III outlines our approach for learning to transfer human grasps to robot grasps and Section IV describes the imitation learning framework. Section V presents the experimental results. Section VI concludes the paper.

II. Related Work

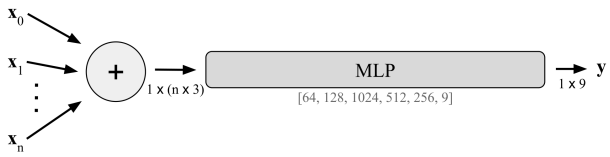
Learning from demonstration or imitation learning is a paradigm in which a robot learns skills by imitating the actions shown by an expert [18] [19]. The approach is popular because it enables robots to learn complex tasks that are otherwise difficult to program. In the robotics context, recent work has shown successful learning of highly advanced skills such as dispensing water from a thermos [20], making coffee [21], preparing a food platter [22] and transferring small items with a kitchen ladle [23].

Human expertise is also a fruitful source of knowledge for learning robotic grasping. This is especially useful for robots operating man-made objects designed for human manipulation. Human knowledge is exploited by physically moving the robot arm and kinesthetic teaching [24], controlling the robot by teleoperation [25] or virtual reality [26], or using a hand-held replica of the robot end-effector [27]. It is cumbersome and time consuming to annotate or physically interact with hardware, therefore, learning from observation [28] is more appropriate because the human involvement is kept to a minimum. But despite this advantage, learning from observation introduces other issues. Most prominent is how to track the human arm and hand while the grasp is performed and secondly how to overcome the disparity between the human and robot hand kinematics. For tracking the hand, many approaches use data gloves or markers with a motion capture system [29] [30] but this is inflexible because the tracking apparatus must be set up and calibrated. This prevents the easy and quick use of the systems by people in real-world home or office settings. Some approaches directly use vision and thus do not require extra hardware to learn and transfer grasps. Do et al. [31] use a single camera to estimate the joints of the human body but simplify the task of estimating the state of the hand. They use a proxy in which the orientation and grasp type are estimated to predict a robot grasp. Therefore, they do not estimate the full hand pose. Palli et al. [32] estimate the wrist and finger tip locations to transfer the demonstrated human grasp to the DEXMART anthropomorphic hand. Due to the high kinematic similarity, the transfer is simplified; it only requires an additional scaling factor between the length of the demonstrator's fingers and the robot fingers. The current solution to transfer grasps to robot end-effectors with significantly different kinematics is to use a predefined mapping between known human and robotic grasp shapes [29]. No work thus far learns a mapping between arbitrary human grasp poses and robotic grasps for grippers with fewer DoF.

Semantic grasping is a special case of grasping in which the grasp enables task-related manipulation [8]. The most common approach is to compute grasps and then introduce constraints or affordances to select grasps that satisfy the task [11]–[13]. Learning semantic grasps directly from observation has been studied in [15]–[17] but these transfer the demonstrated grasps to anthropomorphic hands and some use data gloves to localise hand joints. In this work, we address the problem of observing and imitating semantic grasps only using a camera and provide a learning-based solution to map the human hand to a low-DoF robotic gripper configuration.



(a) Each point passes through a separate MLP (all with shared weights). Feature maps transformed to a global feature with a pooling operation (*maximum* or *average*). Global feature passes through another MLP.



(b) Input points are sorted and concatenated. The concatenated feature vector passes through a single MLP.

Figure 2. Network architectures for regressing robot grasp from human grasp. Layer sizes are shown beneath the MLP blocks.

III. Learning Robotic Grasp Poses from Human Hand Configurations

Imitating a human demonstrated grasp by a robot equipped with a parallel-jaw gripper requires the mapping between the human hand and the gripper's degrees of freedom to be determined. This mapping is represented as a function $\mathcal{F}$ that transforms a human grasp $\mathbf{H} \in \mathbb{R}^H$ to a robot grasp $\mathbf{G} \in \mathbb{R}^G$, i.e., $\mathbf{G} = \mathcal{F}(\mathbf{H})$, and where $H > G$. We choose to model this function as a neural network. The architecture for the network, the loss function and the training procedure are discussed in the following.

A. Network Architecture

An overview of the network architecture is shown in Figure 2a. This architecture is based on the PointNet architecture [33] that is developed for classification or point-wise segmentation of unstructured 3D point cloud data. We modify PointNet to instead regress a 6-DoF pose that represents a grasp for a parallel-jaw gripper. PointNet also includes a spatial transformer network that makes the learned representations invariant to geometric transformations. This is necessary for classification since transformations of the points should not result in different class predictions. For our work, different poses in the input should generate different poses of the grasp, therefore, the spatial transformer network is removed.

The input to the network is a set of points representing the joints of the hand in the camera coordinate system. The joint coordinates are shifted and scaled to fit in the unit sphere. They are then fed to a Multilayer Perceptron (MLP) with layer output sizes $[64,\ 128,\ 1024]$. The output is max-pooled to create a global feature descriptor with 1024 units. The global feature is passed to the second stage of the pipeline to generate the output. This is an MLP that progressively reduces the global feature to the desired size. Our goal is to estimate the 6-DoF pose for a gripper, i.e., translation and rotation. The

 ISBN: 978-1-61208-787-0

translation is represented by the x, y and z coordinates of the centroid of the grasp pose. The rotation is represented as unit vectors of the approach and closing directions. The output, therefore, consists of nine values by passing through the MLP with layer output sizes $[512,\ 256,\ 9]$. Batch norm and ReLu activation function are used for all layers except the last that uses linear activation. Dropout with a rate 0.7 is applied to the second last layer (i.e., layer before the pose prediction).

The output $\mathbf{y} = [\mathbf{t}, \mathbf{a}, \mathbf{c}]$ represents the robot grasp pose. The first three components $\mathbf{t} = (x, y, z)$ is the translation of the gripper with respect to the centre of the hand. The next three, $\mathbf{a} = (a_x, a_y, a_z)$ where $||\mathbf{a}|| = 1$, represent the grasp approach direction as a unit vector. The last three, $\mathbf{c} = (c_x, c_y, c_z)$ where $||\mathbf{c}|| = 1$, represent the closing angle of the gripper as a unit vector. The rotation of the gripper pose is obtained by

$$\mathbf{R} = \left[(\mathbf{c} \times \mathbf{a})^T, \mathbf{c}^T, \mathbf{a}^T \right], \tag{1}$$

and the final transformation matrix is thus $\mathbf{G} = [\mathbf{R}|\mathbf{t}^T]$. This represents the transformation of the gripper to the grasp pose in the camera coordinate system.

Since the input for the regression is a consistent configuration of finger joints, it is in fact unnecessary to account for unordered input with a symmetric function (i.e., pooling operation). Therefore, we also investigate a simplified network, as shown in Figure 2b. The joint values are concatenated and processed by a single MLP with layer output sizes $[64, 128, 1024, 512, 256, 9]$ to be consistent with the baseline approach.

B. Loss

For regression, the l_2 loss is used between the vectors of the estimated, $\mathbf{y}$, and the ground truth, $\mathbf{y}^{\text{GT}}$, gripper poses according to

$$\mathcal{L}\left(\mathbf{y}^{\text{GT}}, \mathbf{y}\right) = \frac{1}{n} \sum_{i=1}^{n} \left(\mathbf{y}_i^{\text{GT}} - \mathbf{y}_i\right)^2. \tag{2}$$

A grasp pose for a parallel-jaw gripper is 180° symmetric around the axis of the approach direction. This has two implications. First, grasps rotated by 180° around the axis should not be penalised but treated the same. Second, annotation does not need to be perfectly consistent for this degree of freedom. To account for the symmetry, the loss is computed for the output $\mathbf{y}$, as well as the same output with the negative of the closing angle, i.e., $\mathbf{y}^{\text{flipped}} = [\mathbf{t}, \mathbf{a}, -\mathbf{c}]$. The symmetric loss function is thus the minimum of the two,

$$\mathcal{L}_{\text{sym}}\left(\mathbf{y}^{\text{GT}}, \mathbf{y}\right) = \min\left(\mathcal{L}\left(\mathbf{y}^{\text{GT}}, \mathbf{y}\right), \mathcal{L}\left(\mathbf{y}^{\text{GT}}, \mathbf{y}^{\text{flipped}}\right)\right). \tag{3}$$

C. Training

The HO-3D dataset [34] is used to train the network. This dataset consists of multiple sequences of people manipulating an object in their right hand. The dataset consists of many different subjects and objects, as well as different perspectives from multiple cameras. The objects used for the data collection are a subset of the YCB object set [35]. The joints of the hand and the pose of the objects are accurately annotated using a joint optimisation procedure (see [34] for more details). A sample of the dataset is shown in Figure 3.

To learn the robot gripper pose corresponding to hand configurations, gripper poses are annotated for the corresponding hand pose, as shown in Figure 3. The transformed object model and hand mesh, as well as the gripper model are loaded into Blender. The gripper model is manually adjusted to align its centre with the wrist position and its direction to approximate the angle between the thumb and the other fingers of the human hand. Fine adjustments are made such that the closure of the gripper tips coincides with the centre of the human grasp. Grasp poses from a single camera perspective (i.e., one subject-object pair) are annotated and the poses are transformed to the other camera perspectives.

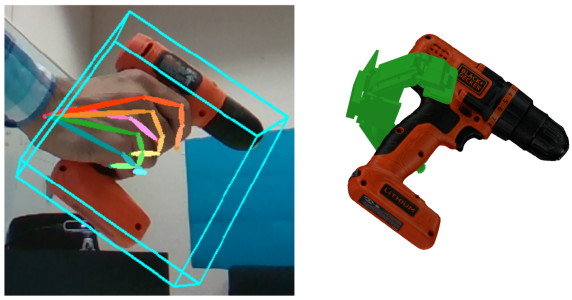


Figure 3. Example of the annotated data for training the gripper pose regression network. Annotation of the hand pose and object from [34] (left). Annotation of the corresponding robot grasp pose (right).

Augmentation is applied to the input. This consists of a global rotation applied to both the input 3D coordinates of the hand joints and to the ground truth pose. The purpose of this augmentation is to generalise the predictions to a larger variety of input pairs. Local augmentation is also applied to the 3D coordinates of the hand joints. This applies both a small rotation, as well as random jitter to individual joints. The purpose of this augmentation is to robustify the network to noisy hand pose estimates.

IV. Imitating Grasp Demonstrations

To imitate grasps with the robot requires the human hand to be estimated online, the hand pose to be translated to a gripper pose and then the gripper pose to be associated with the object of interest. In this work we assume the target objects are known and have a designated local frame of reference. Therefore, every observed grasp is transformed to the local reference and retrieved for new positions of the object.

Pseudo-code for the grasp estimation procedure is given in Algorithm 1. During a demonstration, the pose of the object and the human hand are estimated in each camera frame, I_t, until the pose of the object is observed to move above a set threshold θ_{o}; in other words, until the object is moved by the human demonstrator; or when the object is not detected and thus the pose cannot be estimated due to the occlusion created by the grasp (line 13). This frame, at time t_{end}, establishes the end of the demonstration.

The hand pose at t_{end} is ideal to estimate the robot grasp because it represents the time instant when the human has a solid hold of the object. However, the quality of the estimated pose may be low due to the occlusion that occurs during the physical interaction. Therefore, throughout the demonstration, the hand pose is estimated in every frame (line 7). The nearest frame to the final frame with a valid hand pose, at time t_{est} where $t_{\text{est}} < t_{\text{end}}$, is used to generate the robot grasp pose using the regression network with the relevant hand joints (line 18). The validity of the hand pose depends on the implementation.

 ISBN: 978-1-61208-787-0

Algorithm 1: Robot Grasp Pose Estimation

Result: Robot grasp $\mathbf{G}^o$ in object's frame of reference
1 $\mathbf{P} \leftarrow$ Object pose in first frame
2 $\mathbf{H} \leftarrow$ Computed hand pose
3 $\mathbf{v} \leftarrow$ Grasp pose offset
4 **Loop**
5 Get current image I_t
6 Estimate object pose $\mathbf{P}_t$
7 Estimate hand joints $\mathbf{H}_t$
8 **if** $\mathbf{H}_t$ *is valid* **then**
9 $\mathbf{H} \leftarrow \mathbf{H}_t$
10 **else**
11 $\mathbf{v} \leftarrow$ Update using valid joints in $\mathbf{H}$ and $\mathbf{H}_t$
12 **end**
13 **if** $\mathbf{P}_t = \emptyset$ *or* $|\mathbf{P}_t - \mathbf{P}| > \theta_o$ **then**
14 $\mathbf{P} \leftarrow \mathbf{P}_t$
15 break
16 **end**
17 **EndLoop**
18 $\mathbf{G} \leftarrow$ Estimate robot grasp pose from hand $\mathbf{H}$
19 $\mathbf{G} \leftarrow$ Adjust position of grasp by offset $\mathbf{v}$
20 $\mathbf{G}^o \leftarrow$ Transform $\mathbf{G}$ to object reference frame
21 **Return:** $\mathbf{G}^o$

In this work, we estimate the hand pose using [36] and the residual of the inverse kinematics optimisation, that lifts the 2D keypoints detections to 3D, scores the quality of the hand pose estimate. Please see [36] for more details.

In the case that the hand is only valid in a frame before the final grasp is made (because the hand is less occluded by the object), the predicted gripper pose needs to be adjusted. This is because the hand may be located away from the object and no longer be in an ideal position to associate the robot grasp. To account for this offset, the estimated robotic gripper pose is re-positioned according to the distance and direction between the hand in frames t_{est} and t_{end} (line 19). The vector of the movement in the camera coordinate system for all hand joints between the two frames are computed and then averaged (line 11). For robustness, only the vectors of joints that are visible in every frame between t_{est} and t_{end} are used. This removes spurious estimates that occur due to the occlusion.

Finally, the demonstrated grasp is transformed to the object's frame of reference by $\mathbf{G}^o = \mathbf{P}^{-1}\mathbf{G}$. Once a demonstration is observed, the robot is expected to replicate the grasp for any new pose of the target object. When the object is re-observed, its pose is estimated and the known grasp pose is transformed using the estimate. More concretely, in a new frame, where the object has a different pose $\mathbf{P}'$, the grasp pose that is executed by computing $\mathbf{G}' = \mathbf{P}'\mathbf{G}^o$.

V. Experiments

The performance of the presented method for grasp imitation is evaluated in this section. We first give implementation details. We then report results of offline experiments to quantitatively analyse the robot grasp pose prediction. Lastly, we present results for real-world grasping experiments with a mobile manipulator.

A. Implementation Details

The regression network for robot grasp pose estimation is implemented in PyTorch. All models are trained for 120 epochs with an initial learning rate of 0.001 that is divided by 10 every 50 epochs. A batch size of 64 is used. Training is performed on an NVIDIA GTX TitanX.

Object poses are estimated with Pix2Pose [37], which uses only RGB images as input. The network is trained with the YCB-Video dataset and is therefore compatible with the data from HO-3D. The pose of the hand is estimated using the method in [36] and the keypoints of individual joints to compute the movement offset between the final and valid frame are determined with the OpenCV implementation of [38], where the predictions in the RGB image are lifted to 3D using the corresponding depth image.

Hardware experiments use the Toyota Human Support Robot [39] [40]. Observation of the demonstrations and the stand-alone objects for the grasping experiments use the head-mounted ASUS XTion Pro Live RGB-D camera. The estimated grasp poses are executed by generating a trajectory using MoveIt [41]. This plans trajectories that avoid obstacles within the scene. All code runs on the robot in Ubuntu 16.04 with ROS [42]. Inference for the pose estimation when running on the robot runs on an external PC with an NVIDIA GTX 1050 Ti.

Note that since HO-3D contains examples with the right hand, the applied hand tracking algorithms also use the right hand model and live demonstrations use the right hand. Adapting to the left would require flipping the images in HO-3D and using different models for [36] and [38].

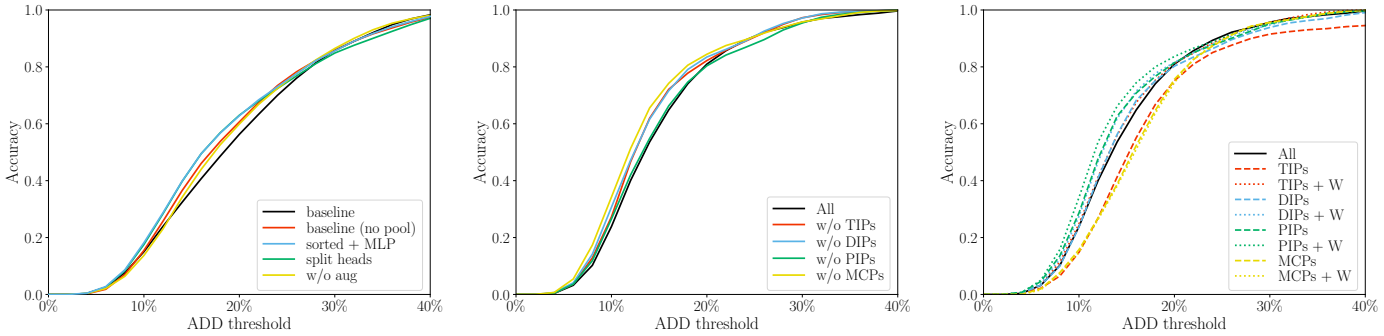
B. Grasp Estimation Analysis

We analyse the quality of the grasp pose estimation using the data from six subjects (ABF, BB, GPMF, GSF, MDF and ShSu) in HO-3D. Separation between the data used for training and testing is maintained by training networks on the data from five subjects and testing on the sixth subject that was not seen in training.

The accuracy of grasp pose predictions is measured by the average distance between all vertices of a 3D mesh (ADD) when transformed by the prediction in comparison to the ground truth. This is a common metric for general object pose estimation [43] because it conveniently unifies translation and rotation error into a single metric. Similar to (3), the accuracy is reported for the minimum of the predicted pose and the 180° rotation around the approach vector to account for the symmetry of the gripper. Points are extracted from the gripper model to compute the metric.

1) Architecture and training procedure: The performance of different architectures and the inclusion of data augmentation is compared in Figure 4a, where results are averaged over all subjects. Considering the different architectures, the best performing is when the joint positions are concatenated into a high-dimensional input that is processed by a single MLP (Figure 2b). This results in 8.8% (at 16% diameter threshold) improvement over the baseline architecture, that applies separate MLPs to each point (Figure 2a). Furthermore, removing the pooling operation and instead processing the the $n \times 1024$ feature vector improves over the baseline architecture. Having separate heads to predict the translation and rotation does not introduce any performance gain, which has been shown in

 ISBN: 978-1-61208-787-0



(a) Architectures and data augmentation.

(b) Missing joint sets.

(c) Only individual joint sets (and wrist).

Figure 4. Grasp pose estimation accuracy at varying ADD thresholds for different architectures, data augmentation and configurations of hand joint inputs.

similar work such as [44]. The data augmentation provides a substantial performance boost. At the 12% diameter threshold, the accuracy drops by 5.5% without data augmentation.

2) Hand configuration: We investigate the relative importance of different joints for learning grasps from human hand poses. The joints are grouped into their corresponding occurrence on the finger. Starting from the ends of the fingers, these groups are the tips (TIPs), distal interphalangeal joints (DIPs), proximal interphalangeal joints (PIPs), metacarpophalangeal joints (MCPs) and the wrist (W). We train three types of networks for each group "X": only with a joint group (X), only with a joint group and the wrist (X + W) and all inputs without the joint group (w/o X). The comparison is conducted for subject ABF (i.e., tested on ABF, trained on all subjects except ABF). The results in Figure 4b show that the removal of a single group of joints in fact improves performance. This is promising for learning grasps from simplified hand poses, especially considering the strong performance without the finger tips, which are often difficult to accurately estimate. In Figure 4c, we show the performance when only specific sets are input to the network. Surprisingly, the performance can be better when only using the DIPs or PIPs. Learning only from the TIPs or MCPs has a noticeable performance loss. This can be explained by the high level of noise in the TIP joint estimates and the relative inflexibility of the MCPs in different hand poses causing a large amount of ambiguity for learning. Including the wrist improves the performance; most notably, the performance of learning from TIPs improves the most when the wrist joint is included and the average performance is slightly better than learning from all 21 joints.

3) Sensitivity to hand pose estimation: The relationship between the robot grasp pose accuracy and the estimate of the hand pose is shown in Figure 5. The hand pose is estimated using [36] and the error of the estimate is computed as the average distance between the estimated joints and the ground truth positions. The robot grasps are predicted from the estimated joints. As the figure shows, there is a clear correlation between the accuracy of the robot grasp and the hand pose. Therefore, the final grasp pose estimate strongly depends on the quality of the input.

C. Real-World Grasp Imitation

Qualitative results of the full grasp imitation pipeline are given in Figure 6. We use the sorted-input single-MLP regression network and train it on all six subjects in HO-3D. The threshold for detecting object movement, θ_o, is set to 5cm.

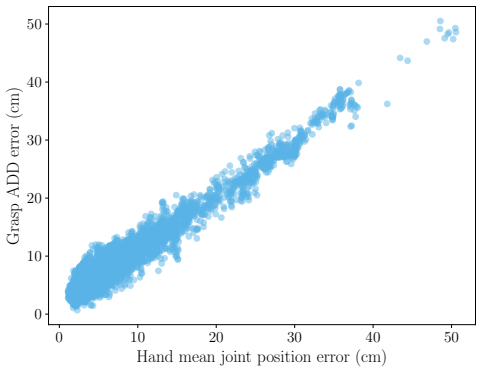


Figure 5. Grasp error (ADD metric) for the estimated hand poses using [36]. Frames in which the hand pose could not be estimated are not included.

TABLE I. GRASP SUCCESS RATE FOR DIFFERENT TARGET OBJECTS FOR THREE DIFFERENT DEMONSTRATIONS. RIGHT-MOST COLUMN SHOWS THE AVERAGE FOR ALL OBJECTS.

	Demo 1	Demo 2	Demo 3	Average
sugar_box	0.6	0.8	0.8	0.73
tomato_soup_can	0.8	0.8	1.0	0.87
mustard_bottle	1.0	0.8	1.0	0.93
mug	1.0	0.4	0.2	0.53
power_drill	0.6	1.0	0.8	0.80

The first column in Figure 6 shows the human demonstration, the second column shows the estimated hand and object pose, the third column shows the estimated robot grasp in the object reference frame, and the remaining columns show successful grasps executed by the robot with the object in different poses. The examples show that the demonstrations of semantic grasps, that is, grasps on the handles of the `mug` (second row) and `power_drill` (third row) are directly transferred to the robot. The robot is able to grasp the relevant part of the object in a similar pose that the person performed.

Table I reports quantitative results of grasps using the full pipeline. For each object, three demonstrations are performed and five grasps are attempted for the object in new poses. The `mug` has the lowest grasp success rate due to the grasps being placed on the thin handle. Consequently, small pose errors cause grasp failures. Surprisingly, the `sugar_box` is also difficult to grasp. This can be attributed to the poor pose estimation that occurred when the front side of the box was not visible. For the other objects, our pipeline achieves a high scoring grasp success rate of over 80%.

 ISBN: 978-1-61208-787-0

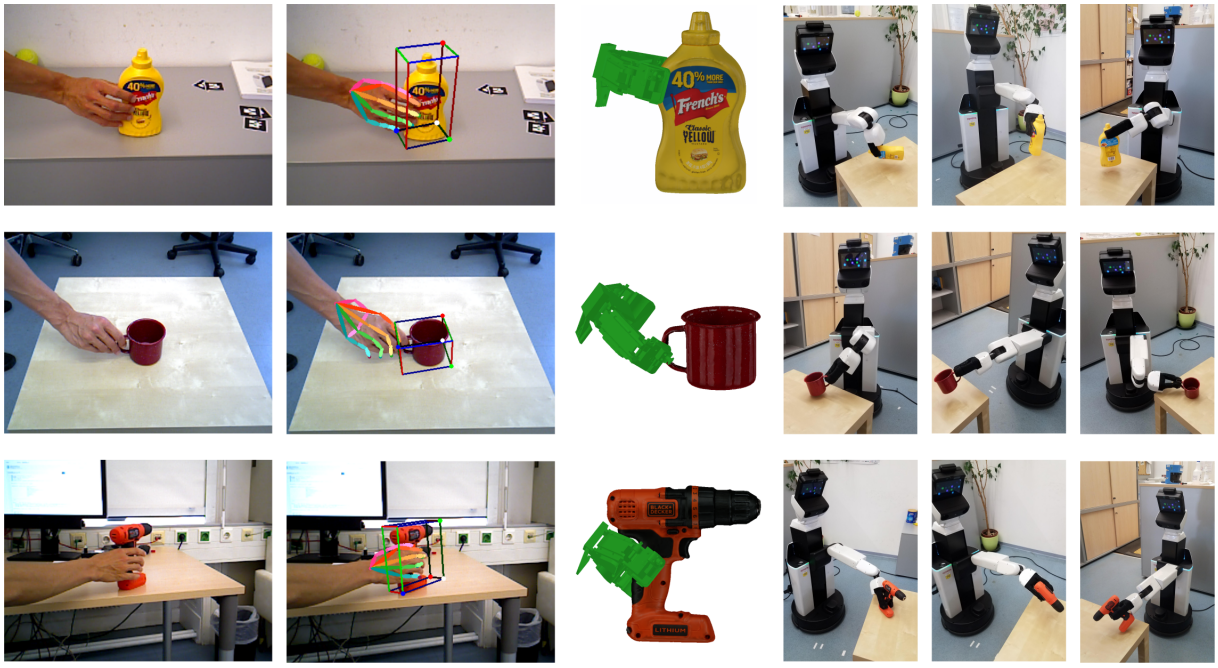


Figure 6. Qualitative results of imitating task-oriented grasp demonstrations for a the `mustard_bottle`, `mug` and `power_drill`. Columns show: (1) observation, (2) estimated object and hand pose, (3) predicted robot grasp, and (4-6) executed grasps for the object presented in new poses.

VI. Conclusion

This work presented an end-to-end system for imitating human-demonstrated task-oriented grasps with a mobile manipulator. Our main contribution is a vision-based imitation learning framework in which the pose of a target object and a demonstrator's hand are tracked to estimate the relevant robot grasp. The robot grasp is derived from the output of a neural network that learns the mapping from a human grasp pose to the configuration of a grasp with a low-DoF gripper. Results show that the predictions of grasps successfully transfer to new demonstrators and that lower-dimensional representations of the hand are sufficient for learning. Experiments with a mobile manipulator demonstrate that a robot is capable of observing a demonstration and immediately grasping the same object when presented in new poses in real-world conditions.

Our analysis showed that the error in the hand pose estimation degrades the quality of the grasp pose. Future work will overcome this issue by directly estimating the robot grasp from the observation without the intermediate hand pose estimation stage. We will also investigate the utility of including an auxilliary task, such as classifying the grasp type, to learn richer features and therefore strengthen the grasp pose estimation task. Lastly, we plan to generalise the framework to transfer grasps from a demonstration to objects that belong to the same class or have similar shape (e.g., grasp all mugs after observing the demonstration for one mug instance) using class-based pose estimation or geometric correspondence prediction.

Acknowledgement

The research leading to these results has received funding from the Austrian Science Fund (FWF) under grant agreement No. I3969-N30 (InDex).

References

[1] Y. Jiang, S. Moseson, and A. Saxena, "Efficient grasping from RGBD images: Learning using a new rectangle representation," in Proc. of IEEE International Conference on Robotics and Automation, 2011, pp. 3304–3311.

[2] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," The International Journal of Robotics Research, vol. 34, no. 4-5, 2015, pp. 705–724.

[3] J. Mahler *et al.*, "Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics," in Proc. of Robotics: Science and Systems, 2017.

[4] A. ten Pas, M. Gualtieri, K. Saenko, and R. Platt, "Grasp pose detection in point clouds," The International Journal of Robotics Research, vol. 36, no. 13-14, 2017, pp. 1455–1473.

[5] C. Goldfeder, M. Ciocarlie, Hao Dang, and P. K. Allen, "The Columbia grasp database," in Proc. of IEEE International Conference on Robotics and Automation, 2009, pp. 1710–1716.

[6] D. Kappler, J. Bohg, and S. Schaal, "Leveraging big data for grasp planning," in Proc. of IEEE International Conference on Robotics and Automation, 2015, pp. 4304–4311.

[7] A. Depierre, E. Dellandréa, and L. Chen, "Jacquard: A large scale dataset for robotic grasp detection," in Proc. of IEEE/RSJ International conference on Intelligent Robots and Systems, 2018, pp. 3511–3516.

[8] Z. Li and S. S. Sastry, "Task-oriented optimal grasping by multifingered robot hands," IEEE Journal on Robotics and Automation, vol. 4, no. 1, 1988, pp. 32–44.

[9] Y. Bekiroglu, D. Song, L. Wang, and D. Kragic, "A probabilistic framework for task-oriented grasp stability assessment," in Proc. of IEEE International Conference on Robotics and Automation, 2013, pp. 3040–3047.

[10] T. Patten, K. Park, and M. Vincze, "DGCM-Net: Dense geometrical correspondence matching network for incremental experience-based robotic grasping," arXiv preprint arXiv:2001.05279, 2020.

[11] R. Detry, J. Papon, and L. Matthies, "Task-oriented grasping with semantic and geometric scene understanding," in Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems, 2017, pp. 3266–3273.

[12] M. Kokic, J. A. Stork, J. A. Haustein, and D. Kragic, "Affordance detection for task-specific grasping using deep learning," in Proc. of

 ISBN: 978-1-61208-787-0

IEEE-RAS International Conference on Humanoid Robotics, 2017, pp. 91–98.

[13] S. H. Kasaei, N. Shafii, L. S. Lopes, and A. M. Tomé, "Interactive open-ended object, affordance and grasp learning for robotic manipulation," in Proc. of IEEE International Conference on Robotics and Automation, 2019, pp. 3747–3753.

[14] K. Fang *et al.*, "Learning task-oriented grasping for tool manipulation from simulated self-supervision," in Proc. of Robotics: Science and Systems, 2018.

[15] J. Aleotti and S. Caselli, "Part-based robot grasp planning from human demonstration," in Proc. of IEEE International Conference on Robotics and Automation, 2011, pp. 4554–4560.

[16] M. Hjelm, C. H. Ek, R. Detry, and D. Kragic, "Learning human priors for task-constrained grasping," in Proc. of International Conference on Computer Vision Systems, 2015, pp. 207–217.

[17] D. Antotsiou, G. Garcia-Hernando, and T.-K. Kim, "Task-oriented hand motion retargeting for dexterous manipulation imitation," in Proc. of European Conference on Computer Vision Workshops, 2018.

[18] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration," Robotics and Autonomous Systems, vol. 57, no. 5, 2009, pp. 469–483.

[19] H. Ravichandar, A. S. Polydoros, S. Chernova, and A. Billard, "Recent advances in robot learning from demonstration," Annual Review of Control, Robotics, and Autonomous Systems, vol. 3, no. 1, 2020, pp. 297–330.

[20] P. N. Hung and T. Yoshimi, "Programming everyday task by demonstration using primitive skills for a manipulator," in Proc. of IEEE 7th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems, 2017, pp. 321–325.

[21] L. Smith, N. Dhawan, M. Zhang, P. Abbeel, and S. Levine, "AVID: Learning multi-stage tasks via pixel-level translation of human videos," arXiv preprint arXiv:1912.04443, 2019.

[22] T. Yu *et al.*, "One-shot imitation from observing humans via domain-adaptive meta-learning," in Proc. of Robotics: Science and Systems, 2018.

[23] Y. Liu, A. Gupta, P. Abbeel, and S. Levine, "Imitation from observation: Learning to imitate behaviors from raw video via context translation," in Proc. of IEEE International Conference on Robotics and Automation, 2018, pp. 1118–1125.

[24] M. Kopicki, R. Detry, M. Adjigble, R. Stolkin, A. Leonardis, and J. L. Wyatt, "One-shot learning and generation of dexterous grasps for novel objects," The International Journal of Robotics Research, vol. 35, no. 8, 2016, pp. 959–976.

[25] T. Zhang *et al.*, "Deep imitation learning for complex manipulation tasks from virtual reality teleoperation," in Proc. of IEEE International Conference on Robotics and Automation, 2018, pp. 5628–5635.

[26] J. S. Dyrstad, E. R. Øye, A. Stahl, and J. R. Mathiassen, "Teaching a robot to grasp real fish by imitation learning from a human supervisor in virtual reality," in Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems, 2018, pp. 7185–7192.

[27] S. Song, A. Zeng, J. Lee, and T. Funkhouser, "Grasping in the wild: Learning 6DoF closed-loop grasping from low-cost demonstrations," IEEE Robotics and Automation Letters, vol. 5, no. 3, 2020, pp. 4978–4985.

[28] V. Krüger, D. L. Herzog, S. Baby, A. Ude, and D. Kragic, "Learning actions from observations," IEEE Robotics Automation Magazine, vol. 17, no. 2, 2010, pp. 30–43.

[29] J. Tegin, S. Ekvall, D. Kragic, J. Wikander, and B. Iliev, "Demonstration-based learning and control for automatic grasping," Intelligent Service Robotics, vol. 2, no. 1, 2009, pp. 23–30.

[30] Y. Lin and Y. Sun, "Robot grasp planning based on demonstrated grasp strategies," The International Journal of Robotics Research, vol. 34, no. 1, 2015, pp. 26–42.

[31] M. Do *et al.*, "Grasp recognition and mapping on humanoid robots," in Proc. of IEEE-RAS International Conference on Humanoid Robots, 2009, pp. 465–471.

[32] G. Palli *et al.*, "The DEXMART hand: Mechatronic design and experimental evaluation of synergy-based control for human-like grasping," The International Journal of Robotics Research, vol. 33, no. 5, 2014, pp. 799–824.

[33] C. R. Qi, H. Su, M. Kaichun, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition, July 2017, pp. 77–85.

[34] S. Hampali, M. Oberweger, M. Rad, and V. Lepetit, "HOnnotate: A method for 3D annotation of hand and object poses," in Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3196–3206.

[35] B. Calli *et al.*, "Yale-CMU-Berkeley dataset for robotic manipulation research," The International Journal of Robotics Research, vol. 36, no. 3, 2017, pp. 261–268.

[36] P. Panteleris, I. Oikonomidis, and A. Argyros, "Using a single RGB frame for real time 3D hand pose estimation in the wild," in Proc. of IEEE Winter Conference on Applications of Computer Vision, 2018, pp. 436–445.

[37] K. Park, T. Patten, and M. Vincze, "Pix2Pose: Pixel-wise coordinate regression of objects for 6D pose estimation," in Proc. of IEEE International Conference on Computer Vision, 2019, pp. 7668–7677.

[38] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand keypoint detection in single images using multiview bootstrapping," in Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017, pp. 4645–4653.

[39] T. Yamamoto, K. Terada, A. Ochiai, F. Saito, Y. Asahara, and K. Murase, "Development of the research platform of a domestic mobile manipulator utilized for international competition and field test," in Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems, 2018, pp. 7675–7682.

[40] T. Yamamoto, K. Terada, A. Ochiai, F. Saito, Y. Asahara, and K. Murase, "Development of human support robot as the research platform of a domestic mobile manipulator," ROBOMECH Journal, vol. 6, no. 4, 2019, pp. 1–15.

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

[42] https://www.ros.org.

[43] S. Hinterstoisser *et al.*, "Model based training, detection and pose estimation of texture-less 3D objects in heavily cluttered scenes," in Computer Vision – ACCV 2012, 2013, pp. 548–562.

[44] G. Gao, M. Lauri, Y. Wang, X. Hu, J. Zhang, and S. Frintrop, "6D object pose regression via supervised learning on point clouds," in Proc. of IEEE International Conference on Robotics and Automation, 2020, pp. 3643–3649.

 ISBN: 978-1-61208-787-0