Few-Shot Object Pose Estimation for Functional Robotic Manipulation

Sebastian Jung[1*], Martin Sundermeyer[2], Maximilian Durner[1], and Rudolph Triebel[1]

[1]German Aerospace Center (DLR)
[2]Google
[*]*Sebastian.Jung@dlr.de*

Abstract

Successfully deploying assistive robots in household environments necessitates the rapid adaptation of the robot's capabilities to functionally interact with newly encountered objects. Existing known-object pose estimation methods encounter scalability challenges due to changing environments, lengthy training times, and reliance on CAD models, hindering quick adaptation to novel objects. On the other hand, approaches working on unknown objects are limited to instance-agnostic manipulations, such as grasping, and lack the ability to perform functional manipulation. In this work, we present the initial results of a general few-shot pose estimation-based approach that seamlessly integrates learning by demonstration with RGB-D templates captured during a scanning and demonstration phase. Our experiments indicate a promising approach that enables a robot to achieve functional manipulations of new objects within an efficient time frame, significantly reducing the time required for adaptation.

1. Introduction

Robots deployed in human environments encounter the challenge of navigating unpredictable scenes that often contain highly dynamic and unfamiliar objects. For successful integration into such settings, robotic systems must exhibit the capability to interact functionally with newly introduced objects, whether they are quasi-static (*e.g.* a new fridge) or dynamic (*e.g.* a mug), without the need for extensive training periods. In this paper, we propose a novel pipeline that not only empowers robots to achieve functional interactions with objects but also enhances the overall assistance provided to humans in their daily lives. Our approach involves a few-shot object pose estimation pipeline, which does not rely on any CAD models, followed by a human-guided demonstration performed by either a caregiver, a family member, or a remote operator. This human-guided demonstration enables the robot to learn complex manipulations, such as operating unfamiliar household appliances, with minimal effort from the user with limited mobility. Our primary focus is on the application of our approach to the assistive wheelchair robot EDAN [14] equipped with a robotic arm, as it operates in various household contexts and encounters novel objects. Imagine a scenario where a person with limited mobility or physical disabilities relies on EDAN's [14] assistance in their daily tasks. However, due to their physical limitations, they might not be able to directly interact with the robot for teaching purposes. Instead, another person or a remote operator can take on the role of teaching EDAN [14] new objects and manipulations through the proposed method. Afterwards, the wheelchair user can instruct the robot to perform the demonstrated manipulations when encountering the previously unknown objects in their daily life without requiring the help of a caretaker. This makes the entire interaction with the robot more inclusive, as it allows individuals with varying physical abilities to leverage the robot's functionalities effectively.

Our objective is the functional 6-DoF manipulation of novel objects, enabling robots to learn object interaction in a seamless and user-friendly way, enhancing their dexterity and adaptability in real-world scenarios.

2. Related work

In recent years, object grasping methods have shown notable progress, enabling robots to pick up unknown objects without extensive re-training [13, 12, 8]. However, a limitation of these methods is their focus on grasping rather than performing more complex and context-specific interactions with objects.

On the other hand, functional manipulation methods have been developed to teach robots intricate manipulations and tasks beyond grasping [7, 5]. These methods, while

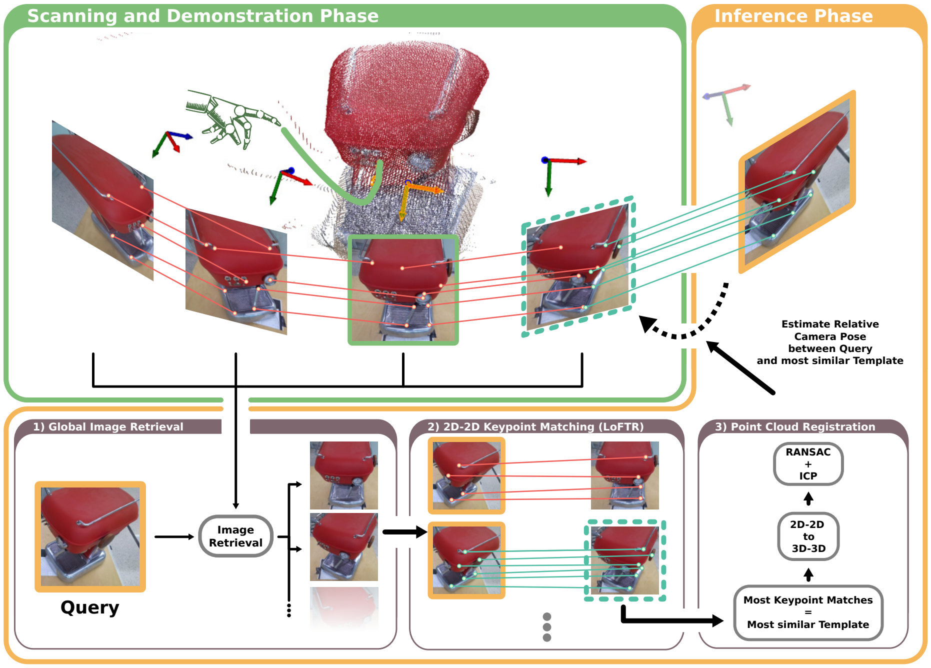
Scanning and Demonstration Phase
Inference Phase
Estimate Relative Camera Pose between Query and most similar Template
1) Global Image Retrieval
Query
Image Retrieval
2) 2D-2D Keypoint Matching (LoFTR)
3) Point Cloud Registration
RANSAC + ICP
2D-2D to 3D-3D
Most Keypoint Matches = Most similar Template

Figure 1. Schematic visualization of our few-shot object poses estimation pipeline for functional object manipulation. A red coffee machine is shown as an example target object. The reference template is shown with a green border. Its corresponding task trajectory is indicated as a green line. The query image is shown with an orange border, the most similar template image with a blue dotted border. Keypoint matches are indicated as lines between the images. The unknown transformation between query and most similar template image is shown as a dotted arrow.

effective for specific object categories, often require significant training efforts for each new task and category, which might hinder scalability and adaptability in dynamic environments with novel objects. Florence et al. [4] proposed a method that learns dense object descriptors from multi-view reconstruction, enabling robots to grasp specific points on novel objects after a relatively short training stage. While promising, this approach still necessitates training and may not fully address the challenge of seamless 6-DoF interaction with completely unknown objects.

Our few-shot pose estimation backbone is inspired by visual localization-based pipelines that include keypoint matching on scenes [9] and objects [11]. By employing a combination of generalized object segmentation and image matching, the robot gains the versatility to interact with objects in a context-sensitive manner. This approach allows the robot to learn complex object-specific manipulations from a small number of template images and a single demonstration, offering potential benefits in terms of time and ease of use.

3. Method

Our approach can be applied to wrist- or externally mounted RGB-D cameras and comprises two main phases: The *scanning and demonstration phase* and the *inference phase*. To separate the object of interest from the background and clutter, both phases utilize generalized instance segmentation techniques, specifically Grounding DINO [6] and MobileSAM [15]. Grounding DINO [6] detects a bounding box based on a textual descriptions of the target object. The bounding box is then used as input to MobileSAM [15] which retrieves a segmentation of the object. This segmentation is then used to crop a square region around the object. As indicated by our ablation study, this approach enhances the robustness of subsequent modules by focusing on the object.

During the scanning and demonstration phase an instructor sequentially captures RGB-D *template* images from different viewpoints using the described combination of Grounding DINO [6] and MobileSAM [15]. Using the mask provided by MobileSAM [15], keypoint matches (LoFTR [10]) between each new template image and its predecessor are determined. RANSAC [3] + Iterative Closest Point (ICP) [2] are then used on the equivalent 3D-3D matches of the point clouds to determine the relative camera poses. An instructor then demonstrates the intended manipulation of the object by guiding the robot arm through an input device or zero-gravity mode, while the 6-DoF end-effector trajectory is recorded. The last template captured before teaching the trajectory is known as the *reference* template view. Optionally, additional template views may be captured that cover object viewpoints expected during inference.

The inference phase can be split into three steps. As a first step, when encountering a previously scanned object from an unknown viewpoint, a crop of the object is received using Grounding DINO [6] and MobileSAM [15] as before. The resulting crop is defined as the *query* image. The query image is compared to all recorded templates using NetVLAD [1] features to retrieve the k most similar templates. As a second step, keypoint matching (LoFTR [10]) is performed between the query image and the retrieved k most similar images. As a last step, the viewpoint between the query image and the template image with the most keypoint matches is estimated using RANSAC [3] + ICP [2] as previously described. Given the relative transformations between the template views calculated during the scanning phase, the viewpoint change between the query and the reference template image in which a task was demonstrated is computed. This enables the robot to repeat the trajectory from a novel viewpoint.

A schematic visualization of our method can be found in Figure 1.

4. Evaluation

Two initial experiments have been performed so far to test our proposed approach. To gain more insight regarding the effect of the object segmentation, we compare the task execution success with and without masking the object using MobileSAM [15] and Grounding Dino [6](see Table 1).

4.1. Experimental Setup

Without masking, we extract the largest possible square crop from the center of the camera image. On the other hand, with masking, we crop a square region, equal to the size of the mask, around the object. Subsequently, the mask is utilized to filter the keypoint matches obtained from LoFTR [10]. Query and template images either both have masks or have no masks. In both cases, we only consider one template image. Two setups are tested: Query and template image have a) the same or b) a different background. For each experiment, the results of the successful execution of a task are recorded for 7 different object rotations and fixed camera positions. Given the positive effect of the masks, in the following experiments, we always consider the object mask.

A total of four different objects were employed in this study, with each object used to teach the robot a specific manipulation task. The robot's execution of these tasks was tested under various settings. For the assessment of our approach, we considered both quasi-static, larger objects, exemplified by opening the *drawer* of a sideboard and flipping a switch of a *coffee machine*, as well as dynamic, smaller objects, demonstrated by pressing the switch of a *multi-plug* and the enter button on a *computer keyboard*. A task execution was deemed successful when the desired goal was achieved. We estimated the maximum translation error in the trajectory to be below $0.5\,\mathrm{cm}$ for successful task completion.

While the current experiments were all performed on a mock-up robotic platform, we plan to conduct additional experiments on a mobile robotic platform. This includes more complex manipulation tasks of dynamic and quasi-static objects in a real household environment. Additionally, we scheduled experiments for quantitative results on vision benchmarks.

4.2. Results

The results of the ablation study in Table 1 underscore the significance of masking an object-centric crop for successful task execution. Notably, only one setup without masking leads to successful execution, and here the query and template images are essentially identical images with the same background and a relative object rotation of $0°$. When the background differs between the query and template images, the performance of LoFTR [10] in focusing on the object deteriorates, even when the object is not rotated relative to the template image. In contrast, the incorporation of a mask enables successful task execution at object rotations of up to $45°$.

Additionally, Table 2 reveals that a higher number of template images increases the maximum object rotation angle under which tasks can still be performed successfully. With the utilization of five templates from different viewpoints around the object, successful task execution can be observed at up to $90°$ object rotation. Conversely, when using only one template, the appearance difference between the query image and the template image is too large at higher rotations, leading to insufficient keypoint matches. This limitation is successfully mitigated by utilizing multiple template images and propagating the 6-DoF trajectory transformations.

4.3. Discussion

Our preliminary experimental results indicate the practicality of our approach. The required number of templates depends mainly on the expected viewpoint variance at inference time. While the manipulation of quasi-static objects such as a fridge only requires a single template, a smaller object with less features and higher viewpoint variance at inference time necessitates a higher number of templates and corresponding masks. Due to our hierarchical matching and the simple scanning and demonstration process, the computational load as well as the demonstration time increase only moderately with more templates. As a result,

BG	Obj	Mask	-45^o	-30^o	-15^o	0^o	15^o	30^o	45^o
Same	Coffee Machine	w/o				✓			
		w		✓	✓	✓	✓	✓	
	Keyboard	w/o				✓			
		w	✓	✓	✓	✓	✓	✓	
Diff	Coffee Machine	w/o							
		w		✓	✓	✓	✓		
	Keyboard	w/o							
		w	✓	✓	✓	✓	✓	✓	

Table 1. Manipulation success of our method under different object rotations using one template view with (w) and without (w/o) object mask, on same and different (diff) background (BG).

Obj	#T	-90^o	-60^o	-30^o	0^o	30^o	60^o	90^o
Drawer	1			✓	✓	✓		
	5	✓	✓	✓	✓	✓	✓	✓
Coffee Machine	1			✓	✓	✓	✓	
	5		✓	✓	✓	✓		✓
Multi-Plug	1			✓	✓	✓	✓	
	5		✓	✓	✓	✓	✓	✓
Keyboard	1		✓	✓	✓	✓	✓	
	5	✓	✓	✓	✓	✓	✓	✓

Table 2. Manipulation success of our method under different object rotations using different amount of template views (#T)

new object manipulations can usually be taught in less than one minute.

4.3.1 Limitations and Future Work

One limitation of our approach is that it requires depth data which can be missing on reflective or black surfaces. However, during the sparse matching process invalid depth pixels can be simply filtered out and partially valid depth maps on objects are often sufficient for accurate, relative pose estimation. Another limitation is that we have not yet integrated a motion planner that takes into account the partially scanned object and the static or dynamic environment. Therefore, transformed end-effector trajectories can lead to collisions especially if object poses change drastically. In future, we also want to train object-centric feature matchers that promise better performance than keypoint matching trained on scenes.

5. Conclusion

Our experiments performed so far show a promising path to teach robotic systems the functional manipulation of newly encountered objects in short time intervals. Specifically, we showed the successful execution of common household tasks from a single demonstration at novel viewpoints of up to 90° rotation. A key insight of our work is the significance of recently introduced generalized object segmentation methods that strongly increase the object viewpoint matching robustness in our robotic experiments. While those initial experiments on four diverse objects show promising qualitative results, more quantitative experiments of our few-shot pose estimation-based method are still required. Furthermore, we are planning to test our pipeline on the EDAN [14] robot in real-world assistive scenarios.

References

[1] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2016. 3

[2] K. S. Arun, T. S. Huang, and S. D. Blostein. Least-squares fitting of two 3-d point sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1987. 2, 3

[3] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Comms. of the ACM*, 1981. 2, 3

[4] P. R. Florence, L. Manuelli, and R. Tedrake. Dense object nets: Learning dense visual object descriptors by and for robotic manipulation. *arXiv:1806.08756*, 2018. 2

[5] S. H. Kasaei, M. Oliveira, G. H. Lim, L. S. Lopes, and A. M. Tomé. Towards lifelong assistive robotics: A tight coupling between object perception and manipulation. *Neurocomputing*, 2018. 1

[6] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv:2303.05499*, 2023. 2, 3

[7] L. Manuelli, W. Gao, P. Florence, and R. Tedrake. kpam: Keypoint affordances for category-level robotic manipulation. In *Int. Symp. of Robotics Research*, 2019. 1

[8] Y. Qin, R. Chen, H. Zhu, M. Song, J. Xu, and H. Su. S4g: Amodal single-view single-shot se (3) grasp detection in cluttered scenes. In *Conference on Robot Learning (CORL)*, 2020. 1

[9] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. *CoRR*, 2018. 2

[10] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou. LoFTR: Detector-free local feature matching with transformers. In *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2021. 2, 3

[11] J. Sun, Z. Wang, S. Zhang, X. He, H. Zhao, G. Zhang, and X. Zhou. OnePose: One-shot object pose estimation without CAD models. In *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2022. 2

[12] M. Sundermeyer, A. Mousavian, R. Triebel, and D. Fox. Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes. In *IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2021. 1

[13] A. ten Pas, M. Gualtieri, K. Saenko, and R. Platt. Grasp pose detection in point clouds. *The Int. Journal of Robotics Research*, 2017. 1

[14] J. Vogel, A. Hagengruber, M. Iskandar, G. Quere, U. Leipscher, S. Bustamante, A. Dietrich, H. Höppner, D. Leidner,

and A. Albu-Schäffer. EDAN: An emg-controlled daily assistant to help people with physical disabilities. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2020. 1, 4

[15] C. Zhang, D. Han, Y. Qiao, J. U. Kim, S.-H. Bae, S. Lee, and C. S. Hong. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv:2306.14289*, 2023. 2, 3