Proceedings of the Joint Austrian Computer Vision and Robotics Workshop 2020 DOI: 10.3217/978-3-85125-752-6-01

Semi-Automatic Generation of Training Data for Neural Networks for 6D Pose Estimation and Robotic Grasping

Johannes Nikolaus Rauer, Mohamed Aburaia, Wilfried Wöber
FH Technikum Wien
{rauer,aburaia,woeber}@technikum-wien.at

Abstract. *Machine-learning-based approaches for pose estimation are trained using annotated ground-truth data – images showing the object and information of its pose. In this work an approach to semi-automatically generate 6D pose-annotated data, using a movable marker and an articulated robot, is presented. A neural network for pose estimation is trained using datasets varying in size and type. The evaluation shows that small datasets recorded in the target domain and supplemented with augmented images lead to more robust results than larger synthetic datasets. The results demonstrate that a mobile manipulator using the proposed pose-estimation system could be deployed in real-life logistics applications to increase the level of automation.*

1. Introduction

Production facilities have successfully deployed classic fixed-programmed robots since the 1960s. Due to their inability to perceive the environment, such robots have mostly been used in mass production, where a static setup can be assumed [8]. The production industries' move away from mass production towards highly customized goods requires increased flexibility. Deploying mobile manipulators, a combination of mobile and articulated robots, for intra-logistical transport tasks, promises this desired modularity [6]. Since the accuracy achieved by mobile robot navigation is not sufficient to grasp objects, robots need sensors to perceive their surroundings and autonomously detect objects' poses [1]. The most promising approaches for pose estimation are machine-learning-based methods applied to camera data [2]. Deep neural networks are trained using annotated ground-truth data – images showing the object and information of its pose [4]. State-of-the-art methods for creating such data use markers rigidly attached to the objects, which have to be removed in cumbersome post-processing [3], or need human annotators that align 3D models to video-streams [5]. In this work an approach to semi-automatically generate 6D-pose-annotated training data using an articulated robot is presented.

2. Semi-Automatic Data-Generation

As shown in Figure 1 the object is placed in front of the robot and a fiducial marker is put on it in a defined pose. The pose of the marker with respect to the camera is computed from the captured image and used to calculate the pose of the object with respect to the robot's base. The marker is captured from multiple perspectives and the mean pose is calculated to minimize errors of the camera calibration and marker detection. Afterwards the marker is removed (care must be taken that the object is not displaced) and the robot arm moves around the object to capture images and associated object-pose data automatically. In order to make the data also usable for training neural networks for object detection, the object can be rendered in a virtual environment to calculate segmentation masks. The design minimizes the extent of human labor. It is only necessary to place the marker on the object, capture images of it, and remove it again, to enable recording of several thousand training images fully autonomously. Drawbacks are that the process has to be repeated to cover the other half of the orientation space and that the background is static. However, this can be solved by data augmentation.

3. Results & Discussion

Multiple annotated datasets are created using the proposed method and used to train the deep-learning-based 6D pose estimation system DOPE [7]. The annotated training data is split into five equally sized

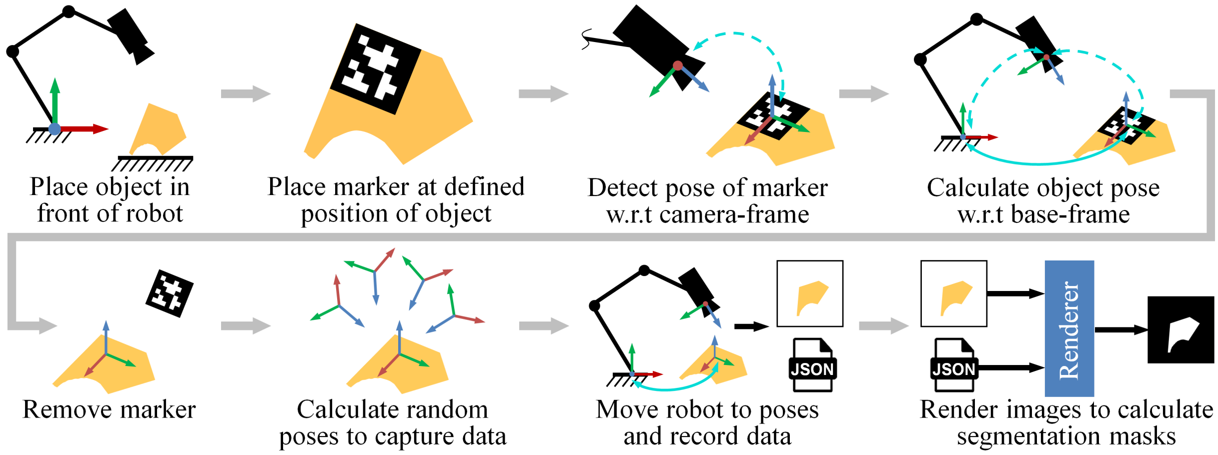
Place object in front of robot
Place marker at defined position of object
Detect pose of marker w.r.t camera-frame
Calculate object pose w.r.t base-frame
Remove marker
Calculate random poses to capture data
Move robot to poses and record data
JSON
JSON
Renderer
Render images to calculate segmentation masks

Figure 1. Procedure for generating annotated data, using a robot and a movable fiducial marker.

portions and merged to gain datasets containing 20% to 100% (15k images) of all recorded samples.

The translational-15mm-error metrics (percentage of tested data for which the translational error is smaller than 15 mm – accuracy necessary for grasping) [7] in Figure 2 show, that using pre-trained models (blue, 6-10) leads to better performance than initializing networks with random weights (red, 1-5). Bigger datasets do not necessarily improve the accuracy since biased datasets lead to wrong generalizations (e.g. network 5). A relatively small dataset recorded in the target domain achieves better results than a several times larger synthetic dataset (network 12: 15k real + 15k domain randomized images), especially when extended using data augmentation (network 11: smallest real dataset augmented twice). The rotational errors show similar results, but are generally lower.

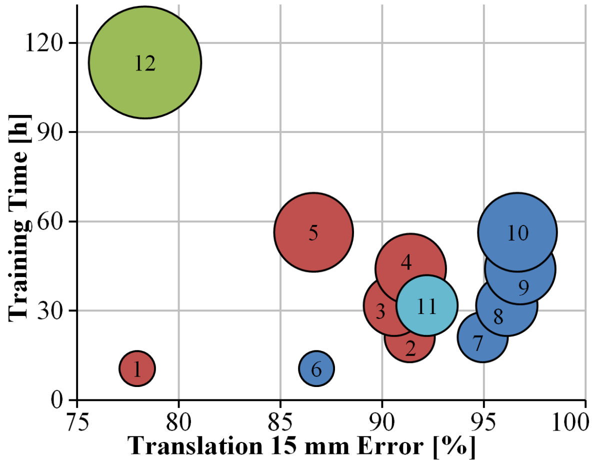

Figure 2. Translational errors compared regarding training time: Synthetic data (green), augmented data (cyan), pre-trained (blue) and non-pre-trained networks (red). Bubble-size visualizes dataset-size.

A qualitative evaluation using a real mobile manipulator confirms that the proposed pose-estimation system could be deployed in real-life logistics applications to increase the level of automation.

References

[1] U. Asif, M. Bennamoun, and F. A. Sohel. RGB-D object recognition and grasp detection using hierarchical cascaded forests. *IEEE Transactions on Robotics*, 33(3):547–564, 2017.

[2] G. Du, K. Wang, and S. Lian. Vision-based robotic grasping from object localization, pose estimation, grasp detection to motion planning: A review. *CoRR*, 2019.

[3] M. Garon, D. Laurendeau, and J. F. Lalonde. A framework for evaluating 6-DOF object trackers. In *15th European Conference on Computer Vision – ECCV*, pages 608–623, 2018.

[4] I. J. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, Cambridge, MA, USA, 2016.

[5] P. Marion, P. R. Florence, L. Manuelli, and R. Tedrake. Label Fusion: A pipeline for generating ground truth labels for real RGBD data of cluttered scenes. In *IEEE International Conference on Robotics and Automation – ICRA*, pages 1–8, 2017.

[6] D. Pavlichenko, G. M. García, S. Koo, and S. Behnke. Kittingbot: A mobile manipulation robot for collaborative kitting in automotive logistics. In *15th International Conference on Intelligent Autonomous Systems – IAS*, pages 849–864, 2018.

[7] J. Tremblay, T. To, B. Sundaralingam, Y. Xiang, D. Fox, and S. Birchfield. Deep object pose estimation for semantic robotic grasping of household objects. In *2nd Annual Conference on Robot Learning – CoRL*, pages 306–316, 2018.

[8] J. Wallén. The history of the industrial robot. *Technical report from Automatic Control at Linköpings universitet*, 2008.