CNNs for State Estimation of Articulated Objects*

Kamil Młodzikowski and Dominik Belter

Institute of Robotics and Machine Intelligence, Poznan University of Technology,
60-965 Poznań, Poland
kamil.mlodzikowski@put.poznan.pl

Abstract. In this paper, we deal with the problem of state estimation of articulated objects during robotic interaction. The robot equipped with an RGB-D camera has to estimate the joint position and rotation of the articulated object when manipulating the object. The problem of accurate state estimation is challenging due to the properties of the RGB-D sensor. The known solutions require some assumptions about the shape of the objects. In this paper, we propose the application of Convolutional Neural Networks to the state estimation of articulated objects from two pairs of RGB-D images.

Keywords: articulated objects · robot perception · deep learning in robotics

1 Introduction

Mobile-manipulating robots working as personal assistants, that help with daily household tasks, should be capable of operating in an unstructured indoor environment. Similarly, robots operating in warehouses, hospitals, or factories share their workspace with humans and should deal with objects in these environments. When interacting with articulated objects, the robots should estimate the state of the object to determine if the interaction is successful (e.g. the robot opened the door).

In this research, we focus on the problem of estimating the configuration of the rotational articulated object from a sequence of RGB-D images. An example scenario is presented in Fig. 1. We propose a system that estimates the rotation axis and the configuration of the joint (angle of rotation) during the interaction. We utilize a pair of RGB-D images from a depth camera that observes the scene with moving objects. We propose the application of a deep neural network to deal with challenges related to this task (occlusions and missing depth data during interaction). Because end-to-end learning is impossible in this case, we suggest three separate neural networks that form a cascade that solves the given problem.

* The work was supported by the National Science Centre, Poland, under research project no UMO-2019/35/D/ST6/03959.

Fig. 1. In the application scenario the robot observes the scene and, from two pairs of images, concludes about the configuration of the articulated object (blue axis).

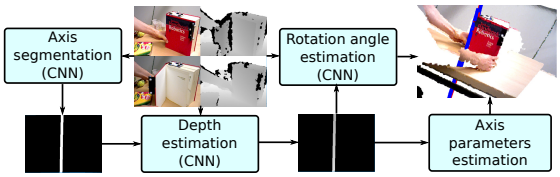

Fig. 2. Block diagram of the procedure for kinematic structure estimation of articulated objects. Each block of the system is explained in detail in the text.

1.1 Related Work

Most of the research on the interaction of robots with articulated objects is focused on visual or force/tactile perception. The RBO dataset [4] contains a set of RGB-D sequences with data from the RGB-D camera, state of the joint, and force/torque data measured during interaction with the objects. In [7] the parameters of the articulated objects, like the axis of rotation/translation or parts' poses are not estimated directly. The knowledge about articulated objects and corresponding actions can be stored in a graph-like structure [3]. The non-parametric belief propagation algorithm proposed in [2] estimates the pose of the articulated objects, but the presented approach assumes that the model of the articulated objects is known in advance. In our research, we applied a set of CNN-based methods that gradually estimate the state of the articulated object. The proposed approach does not require additional assumptions and extracts knowledge directly from training data.

2 Articulated Objects Detection and Estimation

The block diagram of the proposed method is presented in Fig. 2. The first neural network uses two pairs of RGB-D images to perform a segmentation of the rotation axis on an image. The result and the input RGB-D images are passed to the next model. This network estimates the depth value of the previously segmented axis. The third neural network estimates the configuration change of the articulated object (angle of rotation).

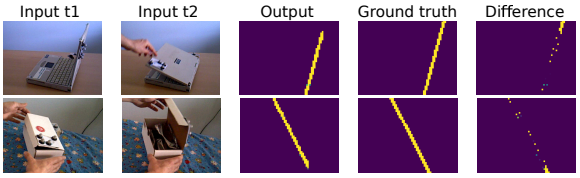

Fig. 3. Example output from the segmentation network compared to ground truth data. *Input t1* is registered at the beginning of the interaction and *Input t2* is the considered frame.

Fig. 4. Example scene visualizations. The red and blue lines represent the ground truth and estimated rotation axes, respectively.

2.1 Axis Segmentation

The CNN that performs the segmentation on the image uses the U-Net with ResNet34 as an encoder trained with the Dice loss. On the input of the CNN, we provide a differential image for the two RGB and depth images and the depth image registered at the beginning of the motion.
Example output of this network is presented in Fig. 3.

2.2 Estimation of the 3D Rotation Axis

The second network estimates a depth image that represents the points on the estimated axis of rotation. This step is needed because the depth from the depth images is not the same as the depth of the rotation axis. The CNN is based on the 3D U-Net [1] with a ResNet3D [6] as a encoder. We propose a new loss function that is a modified MAE loss but is only calculated on the segmented area. The input data is two RGB images, two depth images, and two outputs from the previous network. The output from the network is processed using the RANSAC algorithm, which allows us to extract the best two points that represent the axis in 3D space. Example outputs of the neural network and the obtained axis are presented in Fig. 4.

2.3 Rotation Angle Prediction

The last neural network predicts the joint state change of an articulated object. The used model is the ResNet34 which takes on the input a subtraction of the RGB images, subtraction of depth images, the depth image at the beginning of the motion, and the output from the previous network. We start the training using *Mean Squared Error* as a loss function and then switch to *Mean Absolute Error*.

Table 1. Average error of predicted angle values [rad] for test sequences from the RBO dataset

book10	book22	cardboardbox16	cardboardbox20	microwave12	microwave20	laptop13	laptop17
0.283	0.113	0.096	0.034	0.020	0.028	0.185	0.325

3 Tests and Results

To verify our method, we performed a series of experiments on the test sequences from the RBO Dataset. Taking the first frame of each sequence as a reference, we then iterate through the other frames in the sequence. The results of the tests are presented in Table 1.

4 Conclusion

In this paper, we propose a system that estimates the direction of a rotation axis and angle of rotation from a pair of RGB-D images using a set of neural networks. In contrast to classical methods [5] we do not assume that the surfaces of articulated objects are flat or the robot has a full 3D model of the scene. The average error for the test sequences is 0.135 *rad* (7.8°) which makes the method applicable on real robots. In the future, we are going to integrate the method into our mobile-manipulating robot and verify the system in real-life scenarios.

References

1. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: Learning dense volumetric segmentation from sparse annotation. CoRR **abs/1606.06650** (2016)
2. Desingh, K., Lu, S., Opipari, A., Jenkins, O.C.: Factored pose estimation of articulated objects using efficient nonparametric belief propagation. In: International Conference on Robotics and Automation. Montreal, Canada (2019)
3. Hofer, S., Lang, T., Brock, O.: Extracting kinematic background knowledge from interactions using task-sensitive relational learning. In: International Conference on Robotics and Automation (2014)
4. Martín-Martín, R., Eppner, C., Brock, O.: The rbo dataset of articulated objects and interactions. The International Journal of Robotics Research **38**(9), 1013–1019 (2019)
5. Staszak, R., Molska, M., Młodzikowski, K., Ataman, J., Belter, D.: Kinematic structures estimation on the RGB-D images. In: 2020 25th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA). vol. 1, pp. 675–681 (2020)
6. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. CoRR **abs/1711.11248** (2017), http://arxiv.org/abs/1711.11248
7. Wu, R., Zhao, Y., Mo, K., Guo, Z., Wang, Y., Wu, T., Fan, Q., Chen, X., Guibas, L., Dong, H.: Vat-mart: Learning visual action trajectory proposals for manipulating 3D articulated objects (2021)