TAILOR: Teaching with Active and Incremental Learning for Object Registration

Qianli Xu, Nicolas Gauthier, Wenyu Liang, Fen Fang, Hui Li Tan, Ying Sun, Yan Wu, Liyuan Li, Joo-Hwee Lim

Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore
{qxu, nicolas_gauthier, liang_wenyu, fang_fen, hltan, suny, wuy, lyli, joohwee}@i2r.a-star.edu.sg

Abstract

When a robot is deployed to a new task, it often needs to be trained to detect novel objects. Using deep learning based detector, one has to collect and annotate a large number of images of the novel objects for training, which is labor intensive, time consuming and lack of scalability. We present TAILOR - a method and system for object registration with active and incremental learning. When instructed by a human teacher to register an object, TAILOR is able to automatically select viewpoints to capture informative images by actively exploring viewpoints, and employ a fast incremental learning algorithm to learn new objects without potential forgetting of previously learned objects. We demonstrate the effectiveness of our method with a KUKA robot to learn novel objects used in a real-world gearbox assembly task through natural interactions.

Introduction

In many industrial and domestic applications, automating object detection and recognition is an essential capability. Fast object instance learning becomes a fundamental problem and key requirement of automated systems and machines. Recently, deep learning-based methods have achieved impressive detection accuracy. However, when deployed for a new task, the system needs to be re-trained on new objects, which requires costly human efforts to collect and annotate a large number of images. Some works deal with the generation of high-quality training data by (i) simplifying the data collection and annotation process using specific devices or user interfaces (Marion et al. 2018), or (ii) automating the generation of data samples through augmentation and synthesis (Dwibedi, Misra, and Hebert 2017). For the former, it still involves a lengthy process of data sampling and manual annotation. For the latter, the characteristics of synthesized data need not match with those of actual data from domain-specific environments, leading to unstable and deteriorated performance.

Related work To simplify data collection and annotation, a few works resort to interactive data collection and annotation, where an agent is endowed with the ability to register object instances with human guidance (Kasaei et al. 2015; Dehghan et al. 2019; Kasaei et al. 2018). However,
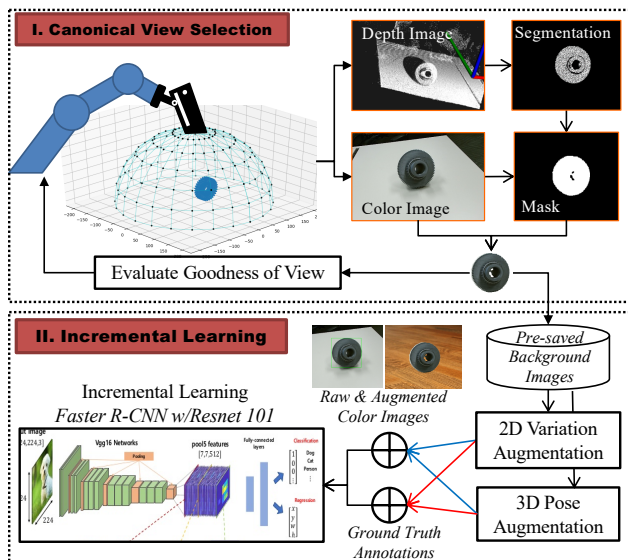
Figure 1: A system architecture of TAILOR
.

they usually require strong prior knowledge of objects (e.g., 3D models) and are restricted by the limited functionalities of the hardware (*e.g.*, robot mobility) and software, thus only addressing small-scale toy problems. It is paramount to fill this gap with an effective interaction protocol for agents to learn new objects in a similar way as a human learner does. Rooted in object detection, object learning can be boosted by using preferential viewpoint information (Xu et al. 2020). For example, the speed and accuracy of object recognition are higher in canonical views relative to non-canonical views, as shown in experiments with both computers and humans (Ghose and Liu 2013; Poggio and Edelman 1990). Here, canonical views refer to viewpoints that visually characterize the entity of interest, which often encompass stable, salient and significant visual features (Blanz, Tar, and Bülthoff 1999).

This paper demonstrates an interactive process facilitated by TAILOR - a method and system that enables *Teaching with Active and Incremental Learning for Object Registration*. Using TAILOR, a robot learns new object instances

while leveraging the representation bias of canonical views. We show how a robot with an eye-in-hand camera interacts with a human teacher to register new objects through (1) human instruction grounding, (2) active image sampling based on canonical view selection, and (3) fast training of an object detector with incremental learning. We present the learning outcome in a real-world use case of interactive object detection. Our method can greatly shorten the cycle-time and reduce manpower cost in new object registration, thus contributing to fast and low-cost deployment of vision-based solutions in real-world scenarios.

Method

The system architecture of TAILOR is shown in Fig. 1. It consists of two phases. In phase I - canonical view selection - the robot uses an eye-in-hand RealSense™camera to get RGB-D images from selected viewpoints, where the viewpoints are evaluated for their goodness of view (GOV) in real-time. The GOV is used to determine canonical views where the RGB images are registered as training samples. In phase II, the selected samples are fed into the training pipeline with a series of data augmentation mechanisms and an incremental learning scheme.

Interaction Protocol

To register new objects, a human user issues speech commands to activate the teaching process. For example, the user puts an object on the tabletop and says *"Start object registration."* Then, the user points at the object and says *"This is the input shaft."*. Through keyword matching and object segmentation, the system registers the object name and relates it with the new incoming training samples. To test the object detector, the user may put one or a few objects on the table and ask *"Where is the input shaft?"*. The robot performs detection to locate the corresponding object and points at it with the gripper.

Canonical View Selection

- **Segmentation** The RealSense™camera provides a color image and point cloud of the scene. Using the point cloud, we adopt the Point Cloud Library (PCL) to estimate the dominant plane model and extract the object on the tabletop. This results in a regional mask of the object, which is used to extract both the color image and depth image within the object region.

- **Goodness of view** The *goodness* of a viewpoint can be measured by multiple metrics defined on different visual features, such as, visible area, silhouette, depth, visual stability, curvature entropy, mesh saliency, *etc.* (Polonsky et al. 2005; Dutagaci, Cheung, and Godil 2010). The GOV is a function of the statistical distributions of visual features that denote its informativeness. To compute GOV, we first perform aforementioned depth segmentation. Next, we use the resultant mask to apply on the color and depth images to extract the regional data of the object. Finally, we compute the GOV based on multiple visual features, including silhouette length, depth distribution, curvature entropy, and color entropy (Xu et al. 2020).

- **Viewpoint exploration** Unlike other interactive learning methods that rely on prior 3D information of an object, our canonical view selection method gathers sparse viewpoint information and evaluates the aggregated GOV of the viewpoints on-the-fly. This denotes a learning experience similar to a robot exploring an unknown object or space. The only assumption is that an RGB-D camera is calibrated with respect to a set of candidate viewpoints on a virtual spherical surface, where an object of interest is located at the sphere center. Thus, the 3D coordinates of the viewpoints are known and the robot can move the camera to those points. The robot then follows the OnLIne Viewpoint Exploration (OLIVE) routine proposed in (Xu et al. 2020) to visit viewpoints. The RGB-D data captured at each viewpoint is used to extract object features and compute the GOV. At any viewpoint, the robot searches the local maxima of GOV, where the GOV is computed as the weighted sum of individual GOV metrics. Once the local maxima is found, the next view is chosen as one with the largest geographical distance to the current view.

For object registration, we put a single object on the table in the current setup. However, this can be extended to multiple objects with two additional components: (1) multiple object segmentation, and (2) hand gesture recognition. For (1), all objects are assumed to be placed in separation and a unique object mask can be extracted for individual objects. With (2) we can extract the pointing gesture so as get the object of interest, i.e., the one being pointed at.

Incremental Learning

The image samples from canonical views are augmented through 2D variation and 3D transformation to generate a training set (Fang et al. 2019). Once the training images of one or a few novel objects have been generated, the system performs on-site incremental learning by calling a recently developed fast incremental learning algorithm for Faster R-CNN (Faster ILOD) (Peng, Zhao, and Lovell 2020). The algorithm uses knowledge distillation for object detectors based on Region Proposal Networks (RPNs) to achieve efficient end-to-end learning. There would be no apparent reduction in accuracy on the newly learned object compared to normal re-training. The incremental learning algorithm is able to avoid catastrophic forgetting of previously learned objects when re-trained just for new objects. The updated model is then applied to detect new objects in new tasks, while still capable of detecting objects learned previously.

Evaluation

We implement TAILOR with a KUKA LBR iiwa 14 R820 robot of 7-degree-of-freedom (7-DoF) and a RealSense™camera mounted on the end-effector. The KUKA robot is a lightweight robot which offers high flexible motion and high-performance servo control with repeatability of ± 0.15 mm. We pre-define a set of 88 viewpoints at the vertices of an icosahedron on a simi-sphere (radius = 350 mm). The geometric center of the object is aligned with the viewpoint sphere center. The camera position is calibrated with a default top view position and its pose is preset so as
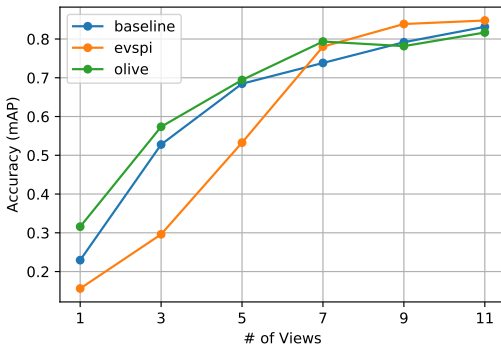
Figure 2: Comparing viewpoint selection on T-LESS dataset
.

to point at the object center at any viewpoint. For a flat object, it is flipped during the registration process so that images of the other side are captured. We perform interactive learning on a few novel components used for the assembly of an industrial gearbox product.

For testing, we put multiple objects (including unknown objects) on the tabletop. The robot is able to reliably detect learned objects and provide corresponding feedback through speech and pointing gestures.

We conducted benchmarking to evaluate the effect of canonical view selection on the T-LESS dataset (Hodan et al. 2017). We compare our method against a baseline random viewpoint selection method and an efficient view selection using proximity information (EVSPI) (Gao, Wang, and Han 2016). The detection performance is shown in Fig. 2. Using only a few image samples, our method is able to train object detectors that outperform alternative methods, especially when the view budget is low. Interested readers may refer to (Xu et al. 2020) for detailed benchmarking results.

Conclusions and Future Work

TAILOR provides an intuitive and efficient way to learn new objects for autonomous agents. It can be used by non-expert users to teach robots new objects without efforts of programming or annotation. As such, the method provides a scalable solution to real-world tasks involving vision-based object detection. To enhance scalability, we intend to simplify the setup by (1) alleviating the need of a dominant flat surface to support the object (which is required for depth-based object segmentation), (2) adopting a mobile robot to allow for more flexible and larger ranges of viewpoint exploration, and (3) employing more powerful natural speech recognition and synthesis (vs. command-type keyword matching) to enhance the human-robot interaction.

Acknowledgments

References

Blanz, V.; Tar, M.; and Bülthoff, H. 1999. What object attributes determine canonical views? *Perception* 28 5: 575–99.

Dehghan, M.; Zhang, Z.; Siam, M.; Jin, J.; Petrich, L.; and Jägersand, M. 2019. Online object and task learning via human robot interaction. *2019 International Conference on Robotics and Automation (ICRA)* 2132–2138.

Dutagaci, H.; Cheung, C.; and Godil, A. 2010. A benchmark for best view selection of 3D objects. In *3DOR '10*.

Dwibedi, D.; Misra, I.; and Hebert, M. 2017. Cut, paste and learn: Surprisingly easy synthesis for instance detection. *2017 IEEE International Conference on Computer Vision (ICCV)* 1310–1319.

Fang, F.; Xu, Q.; Cheng, Y.; Li, L.; Sun, Y.; and Lim, J.-H. 2019. Self-teaching strategy for learning to recognize novel objects in collaborative robots. In *ICRAI'19*.

Gao, T.; Wang, W.; and Han, H. 2016. Efficient view selection by measuring proxy information. *Journal of Visualization and Computer Animation* 27: 351–357.

Ghose, T.; and Liu, Z. 2013. Generalization between canonical and non-canonical views in object recognition. *Journal of Vision* 13 1.

Hodan, T.; Haluza, P.; Obdrzálek, S.; Matas, J. E. S.; Lourakis, M. I. A.; and Zabulis, X. 2017. T-LESS: An RGB-D Dataset for 6D Pose Estimation of Texture-Less Objects. *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)* 880–888.

Kasaei, S.; Oliveira, M.; Lim, G.; Lopes, L.; and Tomé, A. 2015. Interactive open-ended learning for 3d object recognition: An approach and experiments. *Journal of Intelligent & Robotic Systems* 80: 537–553.

Kasaei, S.; Sock, J.; Lopes, L.; Tomé, A.; and Kim, T.-K. 2018. Perceiving, learning, and recognizing 3d objects: An approach to cognitive service robots. In *AAAI*.

Marion, P.; Florence, P.; Manuelli, L.; and Tedrake, R. 2018. Label fusion: A pipeline for generating ground truth labels for real rgbd data of cluttered scenes. *2018 IEEE International Conference on Robotics and Automation (ICRA)* 1–8.

Peng, C.; Zhao, K.; and Lovell, B. 2020. Faster ILOD: Incremental Learning for Object Detectors based on Faster RCNN. *ArXiv* abs/2003.03901.

Poggio, T.; and Edelman, S. 1990. A network that learns to recognize three-dimensional objects. *Nature* 343: 263–266.

Polonsky, O.; Patanè, G.; Biasotti, S.; and Gotsman, C. 2005. What's in an image? Towards the computation of the "best" view of an object.

Xu, Q.; Fang, F.; Gauthier, N.; Li, L.; and Lim, J.-H. 2020. Active image sampling on canonical views for novel object detection. In *ICIP'20*.