ORIGINALARBEIT Elektrotechnik & Informationstechnik (2020) 137/6: 324–330. https://doi.org/10.1007/s00502-020-00817-6

# Learn, detect, and grasp objects in real-world settings

M. Vincze, T. Patten, K. Park, D. Bauer

Experts predict that future robot applications will require safe and predictable operation: robots will need to be able to explain what they are doing to be trusted. To reach this goal, they will need to perceive their environment and its object to better understand the world and the tasks they have to perform. This article gives an overview of present advances with the focus on options to learn, detect, and grasp objects. With the approach of colour and depth (RGB-D) cameras and the advances in AI and deep learning methods, robot vision has been pushed considerably over the last years. We summarise recent results for pose estimation of objects and work on verifying object poses using a digital twin and physics simulation. The idea is that any hypothesis from an object detector and pose estimator is verified leveraging on the continuous advances in deep learning approaches to create object hypotheses. We then show that the object poses are robust enough such that a mobile manipulator can approach the object and grasp it. We intend to indicate that it is now feasible to model, recognise and grasp many objects with good performance, though further work is needed for applications in industrial settings.

Keywords: object detection; learning; recognition; scene understanding; grasping

***Lernen, Erkennen und Greifen von Objekten mit einem mobilen Manipulator.***

*Experten sagen voraus, dass zukünftige Roboteranwendungen einen sicheren und vorhersehbaren Betrieb erfordern werden: Roboter müssen erklären können, was sie tun, um vertrauenswürdig zu sein. Um dieses Ziel zu erreichen, müssen sie ihre Umgebung und die Objekte darin wahrnehmen. Nur dann werden sie die Aufgaben, die sie ausführen müssen, zuverlässig ausführen. Dieser Artikel gibt einen Überblick über aktuelle Fortschritte mit dem Schwerpunkt Lernen, Erkennen und Greifen von Objekten. Mit dem Aufkommen von Farb- und Tiefenkameras (RGB-D) und den Fortschritten bei künstlicher Intelligenz (KI) und Deep-Learning-Methoden wurde die Robotik in den letzten Jahren erheblich vorangetrieben. Es ist bereits möglich, viele Objekte zu modellieren und zu erkennen, obwohl der Nachweis in offenen industriellen Umgebungen noch aussteht. Um dieses Ziel zu erreichen, verwendet man auch die Erkennung größerer Strukturen wie Tische und Wände, um Beziehungen zu den Objekten herzustellen und die Erkennungsraten zu verbessern. Dies wird durch moderne Simulation und digitale Zwillingstechnologie (Digital Twin) unterstützt, mit deren Hilfe überprüft werden kann, ob die erlernten Black-Box-Ergebnisse auch physikalisch Sinn machen. Der Artikel hebt aktuelle Entwicklungen hervor und weist auf zukünftige Trends in Richtung Service- und Industrieroboteranwendungen hin.*

*Schlüsselwörter: Objekterkennung; lernen; Szenen verstehen; greifen*

Received June 16, 2020, accepted July 17, 2020, published online July 30, 2020
© The Author(s) 2020

## 1. Introduction

Ever increasing computing power, availability of large datasets, and modern artificial intelligence (AI) and deep learning approaches have led to new methods in computer vision and robotics. A major influence has been the breakthrough of deep learning, which has enhanced the accuracy and reliability of semantic vision methods such as object recognition and classification [1, 2]. The trends in robotics currently move in two directions. Thanks to new safety features, such as flexible arms, force sensors in joints and touch-sensitive skin, industrial robots are emerging from behind the fences. This creates new mobile applications for the cooperation between robots and humans in manufacturing. And thanks to better AI methods, such as in image processing and robot vision, navigation or path planning, robots are found more and more moving on the workshop floor and in service applications. Examples of this are robots in nursing homes to support the staff, in logistics fulfilling orders, in shops at the reception, and also robots at home that will perform a variety of tasks beyond the vacuuming robots, such as lifting things off the floor [3].

A key aspect to make robots enter all these applications is the understanding of the environment and the objects involved in the tasks. This understanding is essential: the interface to humans will only be possible if the robot shares the interpretation and task at the level of humans, in other words, using the same references and labels to address objects, relations, and actions [4]. Only then it will be possible to solve industrial tasks rapidly and bring robots to services such as making order, Fig. 1.

The contribution of this article is to show that a situated approach increases robustness to the object detection and pose estimation phases of understanding the robot's environment. We first show recent results for pose estimation and then present work on verifying object poses using a digital twin. The idea is that any hypothesis

**Vincze, Markus,** Technische Universität Wien, Automatisierungs- und Regelungstechnik Institute, Wien, Austria (E-mail: markus.vincze@tuwien.ac.at); **Patten, Timothy,** Technische Universität Wien, Automatisierungs- und Regelungstechnik Institute, Wien, Austria; **Park, Kiru,** Technische Universität Wien, Automatisierungs- und Regelungstechnik Institute, Wien, Austria; **Bauer, Dominik,** Technische Universität Wien, Automatisierungs- und Regelungstechnik Institute, Wien, Austria

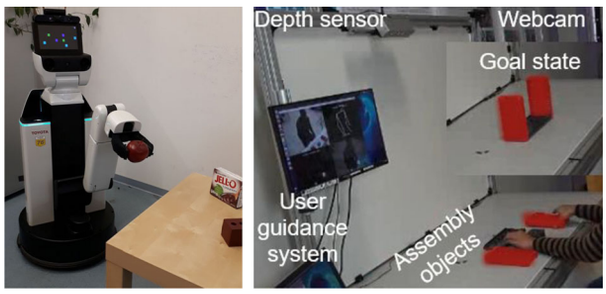



**Fig. 1. Two applications exploiting object recognition and manipulation: Tidying up and a Semantic Assistance System (SAS) to guide object assembly in industrial applications. The idea is that SAS shows the steps to the user to reach the goal state of the part assembly**

from an object detector and pose estimator could be verified with this approach leveraging on the continuous advances in deep learning approaches to create object hypotheses. We then show that the poses are robust enough such that a mobile manipulator can approach the table and grasp the object. Of particular interest is here, that we can use each trial to keep learning which grasps have been successful for improving the reliability of grasping.

The article starts with an outline of the situated approach to robot vision in Sect. 2. We then present our novel approach to object pose estimation, Sect. 3. Section 4 presents the method to verify object hypotheses and Sect. 5 the experiments to use these poses for grasping. Section 6 summarises the results achieves and gives an outlook of future work.

## 2. Situated robot perception: embodied AI from the view of the robot

Classic deep learning methods in computer vision learn from large image datasets created by humans or acquired while driving cars, e.g. [5, 6]. The limit of these approaches is shown when applied to settings in homes and industry [7]. Many of the objects are not recognised. A main reason is that the robot does not create images of target objects in the center under good viewing conditions but is constrained by its motion and physical capabilities, e.g., camera at fixed height, limited viewing angles, and no understanding of back light or good view.

There are two options to counter this shortcoming. One approach, presently pursued in computer vision, creates larger and larger datasets to obtain the view points as gathered when deploying the system [8]. This is the common approach at large companies working on the challenge of autonomous driving, for example, Google[1] or Uber.[2] A similar approach has been taken for the robotics environment to build up a dataset of indoor scenes in [9]. However, this approach is limited, since it will be difficult to create examples of all possible scenarios that one might ever encounter.

Another approach is to work towards better understanding what is actually perceived and done. With respect to perception this means to develop deep learning approaches that use parts and their relationships to better explain overall scenes [10]. With respect to robotics, there is the approach of embodied AI that aims at exploiting the robot body to better understand a situation [11]. We also refer to this approach as situated, since the robot is immersed in its settings and exploits contextual knowledge to improve its understanding of the scenes. Along this approach we develop a situated approach to robot vision that uses contextual knowledge from the map and larger object structures to centre views of the robot and focus object recognition methods [12]. Figure 2 presents an overview. The basic idea is that the robot exploits the data from navigating around for its tasks: (1) floor detection is essential to safely navigate, (2) the boundary of the floor can be perfectly used for localisation, (3) the boundary also clearly delineates areas where there will be larger structures for further analysis such as object recognition, and (4) the larger structures and, in particular, horizontal surfaces are exploited to focus object recognition and classification.

With the detection of larger structures such as cupboards, desks and tables, and the detection of surfaces, the task of detecting objects is enhanced with the dimension to find clusters of data points that stick out of the plane and possibly present one or more objects [9]. This simplifies object detection or can be viewed as presenting a second step of verification to the detection step.

## 3. Object detection and pose estimation

An important task of robots in both industrial and service applications is the detection and tracking of objects. Every handling task of an object requires the recognition of the object and the determination of its position and orientation (object pose). Appropriate methods can be divided into the recognition of objects based on their texture (or appearance) in the colour image or their shape, usually in the depth image. With the advent of RGB-D cameras such as the Kinect (PrimeSense) and RealSense (Intel), the use of colour and depth images has shown to be advantageous for applications in robotics.

The critical task for the eventual goal of grasping an object, is the accurate pose estimation of objects. The inclusion of depth images has induced significant improvements by providing precise 3D pixel coordinates [13]. However, depth images are not always easily available. Hence there is substantial research dedicated to estimating poses of known objects using RGB images only. A large body of work relies on the textured 3D model of an object, which is made by a 3D scanning device, e.g., BigBIRD Object Scanning Rig [14], and provided by a dataset to render synthetic images for training [15]. Thus, the quality of texture in the 3D model should be sufficient to render visually correct images.

The idea of our approach is to convert a 3D model to a coloured coordinate model. Normalized coordinates of each vertex are directly mapped to red, green and blue values in the colour space, see Fig. 3. In this novel method Pix2Pose [16] we predict these coloured images to build a 2D-3D correspondence per pixel directly without any feature matching operation.

An advantage of this approach is that texture information of CAD models are not necessary when limited numbers of real images with pose annotations are available, which simplifies learning for industrial objects, e.g., using texture-less CAD models and collecting images of real objects while tracking camera poses. The method outperforms state-of-the-art methods that use the same amount of real images for training (approximately 200 images per object for LineMOD). In this case, the performance does not rely on the texture quality of 3D models that can be varied with different texturing methods and lighting conditions.

Even though recent studies have shown great potential to estimate poses of multiple objects in RGB images using Convolutional Neural Networks (CNN), e.g., [17], a significant challenge is to estimate correct poses when objects are occluded or symmetric. In these cases the object pose will not be correct or wrongly attached

[1] https://waymo.com/.

[2] https://www.uber.com/at/en/atg/technology/.


© The Author(s)

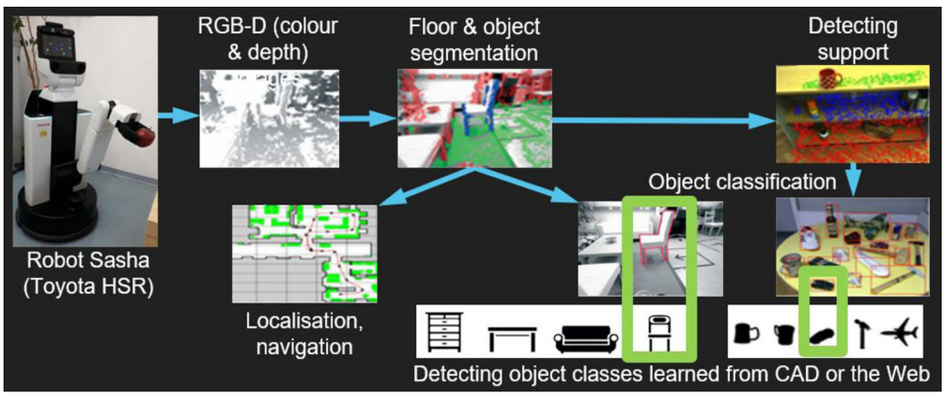



**Fig. 2. Situated object detection exploiting the robot embodiment. Moving through a room, the floor is used for safe navigation, the boundary of the floor for localisation and detecting structures, while structures and features such as planar surface guide object detection**

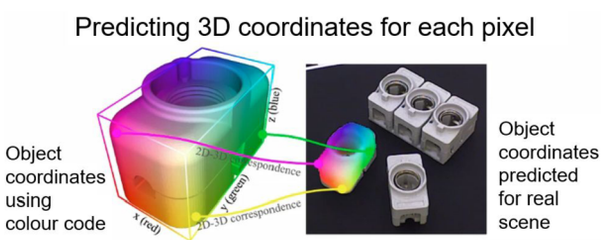



**Fig. 3. Pose estimation based on creating 3D colour coded coordinates, adapted from [16] (Color figure online)**

to a symmetric view resulting in divergence of network training. The ideas presented in Pix2Pose tackle these problems [16]. For occlusion, the pixel-wise prediction is performed for not only the visible area but also the occluded region. For symmetric objects, a novel loss function, Transformer loss, is introduced to guide the predictions to the closest symmetric pose.

The evaluation of pose estimation methods is typically done using several benchmarks, most noticeable, BOP [13]. Two of the benchmarks are specifically related to tasks such as object pose estimation from the view of a robot, the YCB-Video dataset and the Rutgers Amazon Picking Challenge (APC) dataset. At ICCV 2019 [16] won the competition in both these robotics-related benchmarks. Figures 4 and 5 indicate the performance of Pix2Pose when moving around closely placed objects on a small desk. To give an indication of performance, success rate for correct 6D poses on YCB-Video is 75% (up five percent over other methods) with average computing time of 2.9 seconds per scene. APC success rate is 41% (up 24 percent over other methods) using 0.48 seconds per scene. Details are given at the BOP webpage.[3]

## 4. Object hypotheses verification

The idea of a verification step is to exploit the constraints of the robot's environment to check whether an object hypothesis makes sense and is feasible. Object recognition and pose estimation is improved by introducing physical constraints in [18–20]. For example, Aldoma et al. [18] concentrate on a reliable energy function to model the validity of object hypotheses and an efficient method to solve the global model selection problem. Any data point is explained only by one object hypothesis, smooth surfaces should belong to one object, and object and support plane or other objects should not overlap. While these are obvious physical constraints, they are, however, not considered in object recognition methods [1, 2].

The general idea of hypothesis verification is to exploit digital twin technology and physics simulation. Physical object properties

**Fig. 4. Examples for the reliable pose estimation in a real environment. Left, image with original images and bounding box from 2D recognition. Right, the same image with the superimposed object models in 3D. It clearly shows that pose estimates are accurate and cope very well with symmetries**

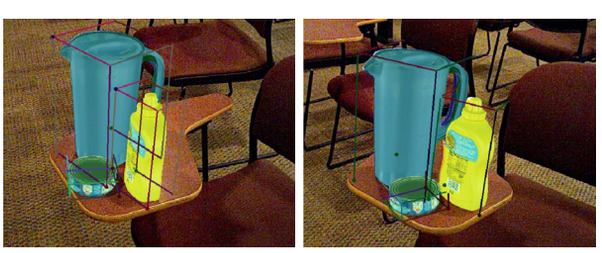


**Fig. 5. Two more examples of 3D pose estimates while moving around the scene in Fig. 4**

[3] https://bop.felk.cvut.cz/.


© The Author(s)

**Fig. 6. Accurate pose estimation using hypotheses verification [23] for images of the Rutgers Extended RGB-D dataset. Particularly in scenes with occluded objects, pose estimation becomes more stable**

are exploited by integrating a physics engine and deep learning. The model predicts physical attributes of objects (i.e., 3D shape, position, mass and friction) using estimates of future scene dynamics from physics simulations. In static scenes, it is highly unlikely to find objects that overcome gravity and rest in unstable positions. This is used in [21] to explain scenes by reasoning about geometry and physics. Especially, the stability of 3D volumetric shapes recovered from 3D point clouds is estimated. Previous approaches focus on predicting the physical behaviour of objects and augmenting hypotheses by modelling Newtonian principles. While Zheng et al. [21] perform physical reasoning utilising mainly depth information, Jia et al. [22] incorporate both colour and depth data.

In our recent work [23] we presented the use of scene models in a feedback cycle to verify the plausibility of hypotheses. The scene model consists of physically annotated objects and structural elements, allowing our approach to inherently consider these properties in a rendering-based verification step using physics simulation. Figure 6 gives examples including partially occluded objects. In addition, this results in an explainable scene description that allows for meaningful interpretation by other system components as well as by humans. The latter is of special significance when deploying autonomous robots in social settings, such as care or educational facilities. For example, objects resting on other objects are clearly identified. In Fig. 6 bottom, right, the glue (white object) rests on the book (red). Consequently, scene explanation reasons, with the help of the physics engine, that taking the book would unsettle the glue.

Based on previous work [18, 24], we investigated how the consideration of physics and appearance cues is integrated into a single, coherent hypotheses verification and refinement framework. As a result, potential false positives in evidence accumulation will be detected as implausible by the hypotheses verification framework. This is in particular helpful when object recognition as well as robot localisation shows typical uncertainty when used in a mobile manipulation scenario. This has been evaluated on the YCB-Video dataset and Fig. 7 highlights some of the results. Success rate of pose estimates with less than 1 cm difference to ground truth, a standard measure in this field, is 91.9 % (up 2.9 percent over other works). For more details, please refer to [23].


**Fig. 7. Samples from the YCB-Video dataset and accurate pose estimates using hypotheses verification**

## 5. Learning to grasp objects

Given accurate object pose, the task of grasping becomes to match the robot motion with the planned pose. While in settings with fixed industrial robots, known objects, and accurate pose this is considered a solved task, grasping from a mobile manipulator adds uncertainty that renders the approach an open challenge for research. Deep learning approaches achieved great results, but for fixed settings: if the input to the vision algorithms varies from the anticipated setting then the task at hand fails, similar to grasp learning in [25], where a disturbance of the camera-robot relation will cause breakdown. Hence, we need an approach that copes with this uncertainty and reacts to it. Ideally, the robot is continuously aware of the situation and acts to obtain data relevant for the task. To date, this is not how robotics is approached in most cases, instead perception systems deliver data at one specific instance triggering a one-shot planning process [26]. Noticeable exceptions are closed-loop control systems such a visual servoing, e.g., [27], and directly learned grasp planners, e.g., [25].

The classical approach to robot grasping is to exploit known objects, e.g., [28, 29]. However, this can only be applied if the set of objects is given and it does not generalise well to new objects. For grasping new objects an option is to use a learned classifier or predictor, e.g., [2, 30] but large amounts of labelled data are required and this is time consuming when done by hand. Another approach is to transfer grasps for known objects to unfamiliar objects. This assumes that a novel object has some similarity to an already learned object and that there is a known successful grasp for the learned object [31]. The idea of these approaches is to exploit a database of sensory observations with associated grasp information, e.g., the object pose and grasp points. The experience is accumulated by trial and error with a robot platform [32] or inferred directly from human behaviour [33]. Grasping an unseen object requires a strategy to map the current observation to the samples in the database and execute (or extrapolate from) the most similar experience. This is typically done using global shape [32, 34], local descriptors [33] or object regions [32]. In contrast to end-to-end learning approaches, experience-based grasping has the potential to learn from very few exemplars.

The key to our incremental approach to grasp learning is to apply a dense geometrical correspondence matching. Familiar objects are identified through global geometric encoding and associated grasps are transferred through local correspondence matching. We


© The Author(s)

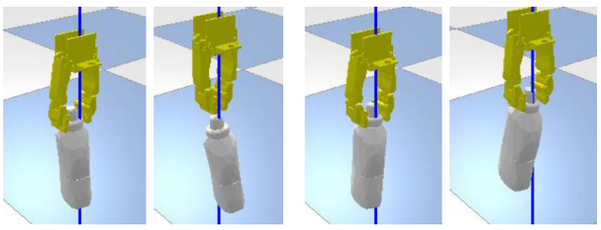


**Fig. 8. Using simulated grasps in PyBullet for learning good and bad grasps. Left two images: example of grasp that failed. Right two images: successful lifting of object**

introduce the dense geometrical correspondence matching network (DGCM-Net) that uses metric learning to encode the global geometry of objects in depth images such that similar geometries are represented nearby in feature space to allow accurate retrieval of experience [35]. The focus is on small objects, where observing the whole object is useful for matching the observation to past experiences. The global geometry is therefore important for the matching task. For the grasp itself, the global geometry is not needed and only the local geometry around the region of interest (i.e., a cube around the grasp points) is used to transfer.

DGCM-Net is used in an incremental grasp learning pipeline, in which a robot self-supervises grasp learning from its own experience. We show that a robot learns to repeatably grasp the same object after one or two successful experiences and also to grasp novel objects that have comparable geometry to a known object. The idea is illustrated in Fig. 8 with two examples from the simulation, a not successful and a successful grasp. Learning which grasps are good or bad helps when transferring to novel objects. The assumption is that previous good grasps have higher likely to also achieve a successful grasp and will be tried first.

We further develop this idea to create generate large quantities of grasp poses by exploiting simulation. Figure 8 illustrates example grasps in the simulator both successful and non-successful. The grasps that succeed in simulation are associated to the relevant objects and then executed with the real robot platform. As shown in Fig. 9, the known successful grasps are detected within the scene even under partial occlusion. To deal with occlusion, we apply data augmentation during training such that samples consist of missing parts. The geometry encoder is guided to generate features that are agnostic to the effects of occlusion. During grasp execution, we use a motion planner to only execute grasps that are reachable. In more detail, our method generates many grasp proposals (as many as there are past experiences) and we use the observed scene to select a grasp that is not in collision [35]. Grasp success rate was 89% as compared to 71% in [36] and 79% in [37].

Figure 9 also shows the top three grasps based on their score. Since the weighting between these two factors is difficult, we rather introduce a ranking. It takes into account a grasp that is as safe as possible and better ranks reliable object poses. Figure 10 shows examples exploiting this method for grasping with a mobile manipulator to demonstrate that given this combined set of methods, grasping objects reliably from a mobile robot is feasible.

## 6. Conclusion

The intention of this article was to highlight that object detection and pose estimation become more and more reliable and move out of the lab to become useful methods for mobile manipulators for a tidy up task or as an assembly assistant (Fig. 1). In particular, we

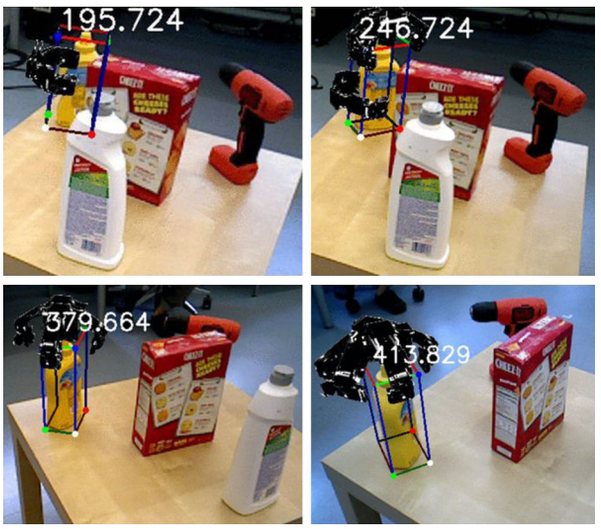



**Fig. 9. Grasp hypotheses while driving around the table shown with an overlay of the black gripper. Note that reachable grasps have been found for partially occluded object in the top row and the change in illumination towards the end of the sequence. The score combines the confidence of the object pose and the clearance to neighboring objects. It is higher for the free-standing objects**

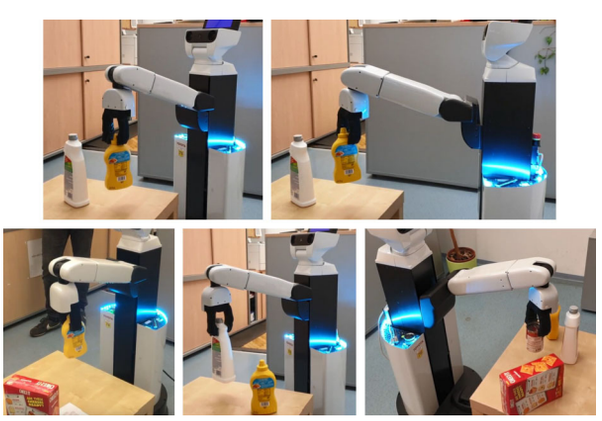


**Fig. 10. Examples of object grasping with the mobile manipulator Toyota HSR**

showed that the situated approach combines contextual information with object recognition methods such that results are more robust. For the task of object pose estimation, the recently developed method Pix2Pose [16] achieved first rank in methods for two challenges (RU-APC and YCB-Video at the pose estimation benchmark at ICCV 2019).

Given the continuous improvement of AI and deep learning methods, it is expected that better object detection and pose estimation methods will appear. Since these methods do not degrade gracefully, we introduced a method for hypotheses verification [23], which uses physics simulation to verify if a pose hypothesis makes sense given the present data and structure of the environment. This also works towards explaining scenes at semantic level. Since pose estimates explicitly consider the support surface, object relations are now transparent and can be retrieved. Furthermore, knowledge from navigating the room adds semantic identity to the surface type, e.g., table, shelf, or counter.


© The Author(s)

Finally, we showed that past experience can be used to incrementally learn grasps for novel objects [35]. This is extended to learn grasps in simulation, which is particularly suitable to industry, where part variations may appear regularly and it is necessary to rapidly, i.e., with a few learning steps, adapt to these parts.

To apply robot vision successfully, other methods such as learning from CAD models [16, 38] as well as learning parts rather than full objects [39] improve recognition results. These results indicate that it becomes feasible that robots begin to tidy up in offices or our homes and execute fetch-and-carry tasks in open environments in industrial settings. The experiments with the Toyota robot in Figs. 9 show that processes run in less than one second without any further optimisation on the on-board laptop resulting in reasonably fluent behaviour for lab demonstrations. However, we see great potential in investigating methods that take the presented results further and optimise the learning of core features and exploit compression to work on limited memory and time resources.

## Acknowledgements

This work was partially supported by the Austrian Science Fund under grant agreement No. I3967-N30 BURG, No. I3968-N30 HEAP, No. I3969-N30 InDex, and the Austrian Research Promotion Agency (FFG) under grant agreement No. 858623 MMAssist.

**Funding Note** Open access funding provided by TU Wien (TUW).

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** Dieser Artikel wird unter der Creative Commons Namensnennung 4.0 International Lizenz veröffentlicht, welche die Nutzung, Vervielfältigung, Bearbeitung, Verbreitung und Wiedergabe in jeglichem Medium und Format erlaubt, sofern Sie den/die ursprünglichen Autor(en) und die Quelle ordnungsgemäß nennen, einen Link zur Creative Commons Lizenz beifügen und angeben, ob Änderungen vorgenommen wurden. Die in diesem Artikel enthaltenen Bilder und sonstiges Drittmaterial unterliegen ebenfalls der genannten Creative Commons Lizenz, sofern sich aus der Abbildungslegende nichts anderes ergibt. Sofern das betreffende Material nicht unter der genannten Creative Commons Lizenz steht und die betreffende Handlung nicht nach gesetzlichen Vorschriften erlaubt ist, ist für die oben aufgeführten Weiterverwendungen des Materials die Einwilligung des jeweiligen Rechteinhabers einzuholen. Weitere Details zur Lizenz entnehmen Sie bitte der Lizenzinformation auf http://creativecommons.org/licenses/by/4.0/deed.de.

### References

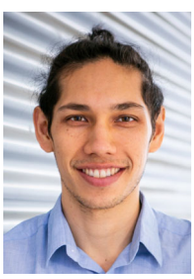
1. Krizhevsky, A., Sutskever, I., Hinton, G. E. (2012): Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (Vol. 25, pp. 1097–1105).
2. Redmon, J., Divvala, S., Girshick, R., Farhadi, A. (2016): You only look once: unified, real-time object detection. In Proc. of IEEE CVPR (pp. 779–788).
3. Fischinger, D., Einramhof, P., Papoutsakis, K., Wohlkinger, W., Mayer, P., Panek, P., Hofmann, S., Körtner, T., Weiss, A., Argyros, A., Vincze, M. (2016): Hobbit, a care robot supporting independent living at home: first prototype and lessons learned. Robot. Auton. Syst., 75, 60–78.
4. Vernon, D., Vincze, M. (2016): Industrial priorities for cognitive robotics. In EUCognition meeting on cognitive robot architectures.
5. Krizhevsky, A., Sutskever, I., Hinton, G. E. (2012): Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, K. Q. Weinberger (Eds.), Advances in neural information processing systems (Vol. 25, pp. 1097–1105). USA: Curran Associates, Inc.
6. Karpathy, A., Joulin, A., Li, F. (2014): Deep fragment embeddings for bidirectional image sentence mapping. http://arxiv.org/abs/1406.5679.
7. Loghmani, M. R., Caputo, B., Vincze, M. (2017): Recognizing objects in-the-wild: Where do we stand? http://arxiv.org/abs/1709.05862.
8. Firman, M. (2016): Rgbd datasets: past, present and future. In 2016 IEEE conference on computer vision and pattern recognition workshops (CVPRW) (pp. 661–673).
9. Suchi, M., Patten, T., Fischinger, D., Vincze, M. (2019): Easylabel: a semi-automatic pixel-wise object annotation tool for creating robotic rgb-d datasets. In Proceeding of the international conference on robotics and automation (ICRA), Montreal, Canada (pp. 6678–6684). https://doi.org/10.1109/ICRA.2019.8793917.
10. Sabour, S., Frosst, N., Hinton, G. E. (2017): Dynamic routing between capsules. http://arxiv.org/abs/1710.09829.
11. Hoffmann, M., Pfeifer, R. Robots as powerful allies for the study of embodied cognition from the bottom up. http://arxiv.org/abs/1801.04819.
12. Potapova, E., Varadarajan, K. M., Richtsfeld, A., Zillich, M., Vincze, M. (2014): Attention-driven object detection and segmentation of cluttered table scenes using 2.5d symmetry. In Proceedings of the 2014 IEEE international conference on robotics and automation (ICRA).
13. Hodan, T., Michel, F., Brachmann, E., Kehl, W., Buch, A. G., Kraft, D., Drost, B., Vidal, J., Ihrke, S., Zabulis, X., Sahin, C., Manhardt, F., Tombari, F., Kim, T., Matas, J., Rother, C. (2018): BOP: benchmark for 6d object pose estimation. http://arxiv.org/abs/1808.08319.
14. Calli, B., Walsman, A., Singh, A., Srinivasa, S., Abbeel, P., Dollar, A. M. (2015): Benchmarking in manipulation research: Using the yale-cmu-berkeley object and model set. IEEE Robot. Autom. Mag., 22(3), 36–52.
15. Sundermeyer, M., Marton, Z., Durner, M., Brucker, M., Triebel, R. (2019): Implicit 3d orientation learning for 6d object detection from RGB images. http://arxiv.org/abs/1902.01275.
16. Park, K., Patten, T., Vincze, M. (2019): Pix2pose: pixel-wise coordinate regression of objects for 6d pose estimation. In Proceeding of the IEEE international conference on computer vision (ICCV), Seoul, South Korea (pp. 7668–7677). https://doi.org/10.1109/ICCV.2019.00776.
17. Tekin, B., Sinha, S. N., Fua, P. (2017): Real-time seamless single shot 6d object pose prediction. http://arxiv.org/abs/1711.08848.
18. Aldoma, A., Tombari, F., Stefano, L. D., Vincze, M. (2016): A global hypothesis verification framework for 3d object recognition in clutter. IEEE Trans. Pattern Anal. Mach. Intell., 38(7), 1383–1396.
19. Mitash, C., Boularias, A., Bekris, K. E. (2018): Improving 6D pose estimation of objects in clutter via physics-aware Monte Carlo tree search. In Proc. of IEEE international conference on robotics and automation (pp. 3331–3338).
20. Zhou, Q. Y., Park, J., Koltun, V. (2018): Open3D: a modern library for 3D data processing. arXiv:1801.09847.
21. Zheng, B., Zhao, Y., Yu, J., Ikeuchi, K., Zhu, S. C. (2015): Scene understanding by reasoning stability and safety. Int. J. Comput. Vis., 112(2), 221–238.
22. Jia, Z., Gallagher, A. C., Saxena, A., Chen, T. (2014): 3D reasoning from blocks to stability. IEEE Trans. Pattern Anal. Mach. Intell., 37(5), 905–918.
23. Bauer, D., Patten, T., Vincze, M. (2020): VeREFINE: Integrating object pose verification with physics-guided iterative refinement. IEEE RA-L Robotics and Automation Letters.
24. Bauer, D., Patten, T., Vincze, M. (2019): Monte Carlo tree search on directed acyclic graphs for object pose verication. In Proceeding of the international conference on computer vision systems (ICVS), Thessaloniki, Greece (pp. 386–396). https://doi.org/10.1007/978-3-030-34995-0-35.
25. Levine, S., Pastor, P., Krizhevsky, A., Ibarz, J., Quillen, D. (2018): Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. Int. J. Robot. Res., 37(4–5), 421–436. https://doi.org/10.1177/0278364917710318.
26. Zhu, Q., Perera, V., Wächter, M., Asfour, T., Veloso, M. (2017): Autonomous narration of humanoid robot kitchen task experience. In 2017 IEEE-RAS 17th international conference on humanoid robotics (humanoids) (pp. 390–397).
27. Bateux, Q., Marchand, E., Leitner, J., Chaumette, F., Corke, P. (2018): Training deep neural networks for visual servoing. In 2018 IEEE international conference on robotics and automation (ICRA) (pp. 3307–3314).
28. Klank, U., Pangercic, D., Rusu, R. B., Beetz, M. (2009): Real-time cad model matching for mobile manipulation and grasping. In 9th IEEE-RAS international conference on humanoid robots (pp. 290–296).
29. Wang, Z., Li, Z., Wang, B., Liu, H. (2016): Robot grasp detection using multimodal deep convolutional neural networks. Adv. Mech. Eng., 8(9), 1–12. https://doi.org/10.1177/1687814016668077.
30. Fischinger, D., Vincze, M., Jiang, Y. (2013): Learning grasps for unknown objects in cluttered scenes. In Proceedings of the international conference on robotics and automation (ICRA) (p. 2013).
31. Bohg, J., Morales, A., Asfour, T., Kragic, D. (2014): Data-driven grasp synthesis – a survey. IEEE Trans. Robot., 30(2), 289–309.
32. Detry, R., Piater, J. (2013): Unsupervised learning of predictive parts for cross-object grasp transfer. In 2013 IEEE/RSJ international conference on intelligent robots and systems (pp. 1720–1727).
33. Liu, C., Fang, B., Sun, F., Li, X., Huang, W. (2019): Learning to grasp familiar objects based on experience and objects' shape affordance. IEEE Trans. Syst. Man Cybern. Syst., 49(12), 2710–2723.
34. Kopicki, M., Detry, R., Adjigble, M., Stolkin, R., Leonardis, A., Wyatt, J. L. (2016): One-shot learning and generation of dexterous grasps for novel objects. Int. J. Robot. Res., 35(8), 959–976. https://doi.org/10.1177/0278364915594244.


© The Author(s)

35. Patten, T., Park, K., Vincze, M. (2020): Dgcm-net: dense geometrical correspondence matching network for incremental experience-based robotic grasping. Frontiers in Robotics, accepted for publication.
36. Fischinger, D., Weiss, A., Vincze, M. (2015): Learning grasps with topographic features. Int. J. Robot. Res., 3, 3.
37. ten Pas, A., Gualtieri, M., Saenko, K., Jr., R. P. (2017): Grasp pose detection in point clouds. http://arxiv.org/abs/1706.09911.
38. Thalhammer, S., Patten, T., Vincze, M. (0002): Sydpose: object detection and pose estimation in cluttered real-world depth images trained using only synthetic data. In Proceeding of the international conference of 3D vison (3DV), Quebec City, Canada (pp. 106–115).
39. Weibel, J. B., Patten, T., Vincze, M. (2020): Addressing the sim2real gap in robotic 3d object classification. In IEEE RA-L (pp. 407–413). https://doi.org/10.1109/LRA.2019.2959497.

## Authors

**Markus Vincze**
is Professor in Robotics at TUW. Presently he leads the "Vision for Robotics" laboratory with the goal to make robots and other machines see. Markus has coordinated four EC projects (RobVision, ActIPret, robots@home, and HOBBIT) and the Austrian Cognitive Vision Network. He was key scientist in EU projects FlexPaint, ECVision, FibreScope, MOVEMENT, XPERO, GRASP and CogX, STRANDS, Squirrel and Flobot. Industrial projects include "Intelligent Teach-In" for Wagner-Biro, interfacing to objects for Aeolus Robotics, Inc., and object class perception and grasping for Omron, Japan. He was program chair of ICRA 2013, general chair of ICVS 2013, organised the European Robotics Forum ERF 2015 and HRI 2017 in Vienna. His special interest are machine vision techniques for real-world robotics solutions.



**Timothy Patten**
is a postdoctoral researcher in the Vision for Robotics laboratory at TUW. He received his PhD from the Australian Centre for Field Robotics at the University of Sydney, Australia. During his PhD he worked on a number of field robotics projects and developed methods for active object classification/recognition, object recognition from LIDAR and state estimation for environment modelling. Since the start of 2017, he has been at TUW, primarily involved with the EU project Squirrel and the industry linked project with Aeolus Robotics, Inc. In this time he has worked closely in object segmentation, recognition, grasping and task planning. Currently, he is the principle investigator at TUW in the CHIST-ERA project InDex, for which he is developing methods for object tracking and semantic grasping.

**Kiru Park**
is a PhD student and research assistant in the Vision for Robotics laboratory at TUW since October of 2016. Before he started his PhD, he worked in Hyundai Motor Company for 5 years as a research engineer in the Ergonomics and Human-Machine Interface team after he received his master's degree from the Department of Mechanical Engineering at KAIST, Korea. During his PhD, he has focused on 6D pose estimation of objects while he primarily involved with the EU project STRANDS and the industrial project linked with OMRON Corporation. Currently, he is developing methods for self-supervised collection and annotation of training images for object detection, segmentation, and pose estimation using robots.

**Dominik Bauer**
is a PhD candidate and research assistant with the Vision for Robotics group at TU Wien. He holds a Master's degree in Visual Computing as well as a Bachelor's degree in Media Informatics and Visual Computing, both from TU Wien.


© The Author(s)