VLRA: Vision and Learning for Robotic Applications

Editorial

Special Track along with ICAS 2020
The Sixteenth International Conference on Autonomic and Autonomous Systems
September 27 - October 01, 2020 - Lisbon, Portugal
https://www.iaria.org/conferences2020/ICAS20.html

Timothy Patten
Automation and Control Institute, TU Wien, Vienna, Austria
patten@acin.tuwien.ac.at

Geraldo Silveira
Robotics and Computer Vision research group, Center for Information Technology Renato Archer (CTI), Campinas-SP, Brazil
geraldo.silveira@cti.gov.br

***Abstract*—Vision and learning algorithms are integral components of robotic systems. Research in these areas receives significant attention and the scientific advances are widespread. However, much work is still only evaluated on datasets or in laboratory conditions. Consequently, the usability for robotic applications in the real world, in general, remains to be truly validated. This special track, "Vision and Learning for Robotic Applications" promotes the development of vision and learning for solving applications with a focus on technological advances for visual sensors; data acquisition and datasets; robot vision and learning-based methods in the wild; interfacing vision and learning; integration for high-speed control; and methods applied to navigation, semantic understanding, manipulation, decision making and cognitive behaviour.**

Keywords–Robot vision; Robotic datasets; Vision and learning in the wild; Object detection, recognition and tracking; Visual learning; High-speed control; Deep learning; Grasping and manipulation; Robots for service, manufacturing, autonomous driving, transportation, agriculture and construction.

I. INTRODUCTION

Computer vision and machine learning algorithms are commonly deployed in robotics for a broad variety of applications such as service, manufacturing, autonomous driving, transportation, agriculture and construction. Whether a robot is indoors, such as in warehouses or homes, or outdoors, such as on roads or farms, visual understanding and learning is integral in order to operate continuously in these real-world environments. While research in vision and learning receives significant attention, much work is still only evaluated on standard datasets or in controlled laboratory conditions. This means that the usability of algorithms for robotic applications in the real world is often untested. As our expectation of robot performance increases and our anticipation of seeing robots in our everyday lives grows, the field is facing a new challenge to develop algorithms that are robust and stable for these scenarios. The outcome of the research is therefore profound and has the potential to be applied to many subfields in robotics such as navigation, semantic understanding, manipulation, decision making and cognitive behaviour.

Similar to other fields of science, robotics has progressed significantly in recent years due to the breakthrough of deep learning. Traditional vision tasks, such as recognition and classification [1], were the first to profoundly change the landscape of current methodologies. Today, we see not only deep learning applied to complex problems but also applied in an end-to-end fashion [2]. Methods are now so powerful that they even surpass human capabilities [3]. It is therefore no surprise that a number of submissions in this special track incorporate deep learning in the robotic system.

Despite these advances, most algorithms are only evaluated in constrained settings and when deployed on a robot in a new setting, the performance is underwhelming. For example, with object recognition, many algorithms sustain a significant performance loss when tested on data collected by a robot as opposed to data captured by humans [4]. This calls for both careful consideration when deploying such algorithms as well as further attention in their development [5]. The purpose of this special track is precisely the promotion of these considerations and developments.

A key aspect for any vision algorithm is the data; therefore, some important considerations are the types of sensors used, the acquisition methods applied, and the datasets generated for both training and testing. These should be taken into account to enable the progress of robot vision into the wild and to overcome the challenges of deploying learning-based methods in real-world systems. Developing such systems provides valuable insight not only about individual algorithmic performance but also the interfaces between algorithms. Furthermore, the experiences are informative and help to better understand the interface between the concepts of vision and learning themselves. Systems should perform at high speeds, not only to be efficient but also to enable reactive behaviour necessary for safe operation. Therefore, integrating vision and learning techniques for high-speed control is a challenging problem.

In this editorial, Section II summarises the contributions to the special track that address one or multiple aspects discussed thus far. Then in Section III, conclusions are drawn and we reflect on future research directions.

II. Special Track Submissions

The first paper, "Imitating Task-oriented Grasps from Human Demonstrations with a Low-DoF Gripper" by Patten and Vincze [6], develops a framework for a robot to learn how to grasp objects by directly observing a human demonstration. This is highly relevant for robots operating in the real world because they will often encounter man-made objects, and as such should handle those objects according to their design. A key component of the approach is a neural network, inspired by the PointNet architecture [7], that translates the higher-DoF human hand pose to the lower-DoF parallel-jaw gripper of the domestic mobile manipulator. This network is trained using an established dataset of human hand poses with new annotations of the corresponding robot gripper poses. In the analysis, the network is shown to have good accuracy and robustness to missing inputs (i.e., sets of hand joints). The data augmentation applied during training reveals a performance boost in the tests. The authors present the developed pipeline with real-world grasping experiments for a selection of objects. Overall, the objects are grasped with a high success rate when presented in new poses after only observing a single human demonstration.

The second paper, "Computation of Suitable Grasp Pose for Usage of Objects Based on Predefined Training and Real Time Pose Estimation" by Chowdhury et al. [8], is similar to the first paper [6] as it also presents an approach for learning task-relevant grasps. Once again, the motivation is established at enabling robots to operate in man-made environments and perform everyday tasks. In this work, the robot arm is positioned with respect to the object of interest and both poses (i.e., of the object and the robot wrist) are recorded. When the object is presented in a new pose, this pose is estimated and the grasp is transformed for the new scenario. This work integrates two object pose estimation techniques for two types of objects: a color-based method for linear objects and a homography-based method for objects with a flat surface but more complex shape. The evaluation uses a PR2 mobile manipulator and demonstrate grasping with both arms. A high grasp success rate is achieved for objects attached to a tripod and also presented in the hand of a human participant.

The third paper, "Semantic Segmentation for the Estimation of Plant and Soil Parameters of Agricultural Machines" by Riegler-Nurscher et al. [9], considers the use of vision algorithms on agricultural machines to improve agronomic process and evaluation through automation. Due to the outdoor and uncontrolled environment, these algorithms are challenging due to occlusion, clutter and illumination variation. The focus of the work is to show that many applications can be addressed by semantic segmentation. To that end, the paper shows that soil cover estimation, grass-legume ratio estimation, grassland swath detection and grassland cut segmentation are suitable tasks for a single algorithm. The ERFNet [10], an encoder-decoder Convolutional Neural Network (CNN) architecture, is fine-tuned using manually annotated data relevant for the specific tasks. The experiments show that better performance is achieved when the network is first pre-trained on a larger dataset, even though it is a different task. For example, pre-training on a dataset of clovers pasted on grass, then fine-tuned on specific tasks is better performing than training the network from scratch with only the task relevant dataset. This is indeed a promising result because combining standard datasets with a small amount of specific data can be a cheap or less time consuming pathway to train a model for a real-world task.

The fourth paper, "Reference Detection for Off-road Self-driving Vehicles using Deep Learning" by Pederiva and de Paiva [11], evaluates the usability of deep learning-based object detectors for localising a landmark object during off-road autonomous driving. The application of autonomous driving is one of the most mature in modern robotics. This work considers the challenging domain of off-road driving (i.e., non-surfaced, uneven terrain). Detecting the reference object, a cone, is useful in this case since it can be used to outline the boundary of drivable terrain. The three architectures evaluated demonstrate different performance and are concluded to be suitable for different conditions. Fast YOLOv2 [12] has the lowest accuracy but also the lowest computation time, therefore, it would be most suitable during tight curves or emergency situations, where fasts decisions need to be made. Faster R-CNN [13] has higher accuracy but at the price of more computation. Therefore, this is more suitable for detecting references further away and in particular on straight paths. MobileNetv2 SSD300 [14][15] is a compromise, having almost the same accuracy as Faster R-CNN with computation twice as fast. The insights gained from the study are highly useful as they indicate in which regimes of operation certain algorithms are more suitable. Furthermore, the tests in the outdoor scenarios provide real-world evidence of their application.

The fifth paper, "In the Depths of Hyponomy: A Step Towards Lifelong Learning" by Boccato et al. [16], introduces a new framework for lifelong learning of semantic classes. This is a desirable capability, as it can extend the operational time of robots that operate in real-world and uncontrolled environments. In this theoretical framework, classes are not fixed but instead the intra-class variability is monitored over time to trigger a refinement of the categories encoded in a classifier. An appropriate metric is proposed to quantify the intra-class variability that ultimately leads to the class restructuring. Experiments conducted with the DeepNCM classifier [17] show that the metric correctly reports less consistency of the deep representations when samples from distinct subclasses are introduced. As such, this constitutes an opportunity to "split" the initial classes into its more distinguishable subclasses. Furthermore, quantitative evidence is provided that the metric accurately identifies classes that should be split or maintained. The results and the framework itself are intriguing because they present a new opportunity for robots to operate longer in the wild. By re-configuring their own classification definition in a fully self-supervised way, there is no need for human intervention. Thus, robots can continue to work, possibly in harsh or remote conditions, and adapt by themselves based on their own experience.

The sixth paper, "Towards a Unified Approach to Homography Estimation using Image Features and Pixel Intensities" by Nogueira et al. [18], proposes a hybrid method for feature and intensity-based homography estimation, which has numerous real-world applications such as in image mosaicing, visual tracking, visual servoing and grasping. Traditional methods use either image features or pixel intensities for the underlying image registration problem. This work, however, combines both approaches in a single optimisation procedure and thus leverages the advantages of the individual ones. The conducted experiments show that the proposed approach has a higher convergence domain than the individual methods. In another set of experiments, the work shows that the unified

approach converges to a lower error value than the feature-based method. This error value is indeed the same final value of the intensity-based approach, however, the unified approach reaches convergence with less processing time. The developed work runs in real-time and is thus highly suitable for the stated applications. In addition, the C++ library and Robot Operating System (ROS) package are publicly available for immediate use on robotic platforms.

III. Conclusion

The VLRA special track has brought together a diverse collection of articles related to vision and learning in robotic systems. The submissions have covered the deployment of algorithms for general robotic tasks but also specific applications such as agriculture and autonomous driving. Indeed, the developments or analyses of grasp estimation, semantic segmentation, object detection, lifelong learning and homography estimation will help the community to advance robot capabilities related to a range of tasks. In the future, we see that vision, learning and their combination will continue to be a core aspect of robotic systems. The contributions of this special track only address a small set of challenges but many more deserve similar treatment. We hope that our promotion of real-world deployment, testing and evaluation is pursued in the future to encourage the migration of robots from the laboratory into the wild.

Acknowledgment

We thank the hard working organisers of ICAS 2020 for their constant support of the VLRA special track. We also thank all the authors for their interesting contributions. Last but not least, we thank the anonymous reviewers who reserved their time to provide valuable feedback that most certainly resulted in improvements and also hopefully sparked fresh ideas for new research endeavours.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems 25, 2012, pp. 1097–1105.

[2] H. A. Pierson and M. S. Gashler, "Deep learning in robotics: A review of recent research," Advanced Robotics, vol. 31, no. 16, 2017, pp. 821–835.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in Proceedings of tje IEEE International Conference on Computer Vision, 2015, pp. 1026–1034.

[4] M. R. Loghmani, B. Caputo, and M. Vincze, "Recognizing objects in-the-wild: Where do we stand?" in Proceedings of the IEEE International Conference on Robotics and Automation, 2018, pp. 2170–2177.

[5] N. Sünderhauf, O. Brock, W. Scheirer, R. Hadsell, D. Fox, J. Leitner, B. Upcroft, P. Abbeel, W. Burgard, M. Milford, and P. Corke, "The limits and potentials of deep learning for robotics," The International Journal of Robotics Research, vol. 37, no. 4-5, 2018, pp. 405–420.

[6] T. Patten and M. Vincze, "Imitating task-oriented grasps from human demonstrations with a low-DoF gripper," in Special Track: Vision and Learning for Robotic Applications (VLRA), along with ICAS, 2020.

[7] C. R. Qi, H. Su, M. Kaichun, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017, pp. 77–85.

[8] M. T. Chowdhury, S. K. Paul, M. Nicolescu, M. Nicolescu, D. Feil-Seifer, and S. Dascalu, "Computation of suitable grasp pose for usage of objects based on predefined training and real time pose estimation," in Special Track: Vision and Learning for Robotic Applications (VLRA), along with ICAS, 2020.

[9] P. Riegler-Nurscher, J. Prankl, and M. Vincze, "Semantic segmentation for the estimation of plant and soil parameters on agricultural machines," in Special Track: Vision and Learning for Robotic Applications (VLRA), along with ICAS, 2020.

[10] E. Romera, J. M. Álvarez, L. M. Bergasa, and R. Arroyo, "Erfnet: Efficient residual factorized convnet for real-time semantic segmentation," IEEE Transactions on Intelligent Transportation Systems, vol. 19, no. 1, 2018, pp. 263–272.

[11] M. E. Pederiva and E. C. de Paiva, "Reference detection for off-road self-driving vehicles using deep learning," in Special Track: Vision and Learning for Robotic Applications (VLRA), along with ICAS, 2020.

[12] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6517–6525.

[13] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, 2017, pp. 1137–1149.

[14] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 4510–4520.

[15] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in Proceedings of the European Conference on Computer Vision, 2016, pp. 21–37.

[16] T. Boccato, T. Patten, M. Vincze, and S. Ghidoni, "In the depths of hyponymy: A step towards lifelong learning," in Special Track: Vision and Learning for Robotic Applications (VLRA), along with ICAS, 2020.

[17] S. Guerriero, B. Caputo, and T. Mensink, "DeepNCM: Deep nearest class mean classifiers," in Proceedings of the International Conference on Learning Representations - Workshop, 2018.

[18] L. Nogueira, E. Paiva, and G. Silveira, "Towards a unified approach to homography estimation using image features and pixel intensities," in Special Track: Vision and Learning for Robotic Applications (VLRA), along with ICAS, 2020.