Warehouse Picking Automation Workshop (WPAW),
IEEE International Conference on Robotics and Automation (ICRA), Singapore, May 2017.

Data-efficient Deep Learning for RGB-D Object Perception in Cluttered Bin Picking

Max Schwarz and Sven Behnke

***Abstract*— Deep learning methods often require large annotated data sets to estimate their high numbers of parameters, which is not practical for many robotic domains. One way to migitate this issue is to transfer features learned on large datasets to related tasks. In this work, we describe the perception system developed for the entry of team NimbRo Picking into the Amazon Picking Challenge 2016. Object detection and semantic Segmentation methods are adapted to the domain, including incorporation of depth measurements. To avoid the need for large training datasets, we make use of pretrained models whenever possible, e.g. CNNs pretrained on ImageNet, and the whole DenseCap captioning pipeline pretrained on the Visual Genome Dataset. Our system performed well at the APC 2016 and reached second and third places for the stow and pick tasks, respectively.**

I. Introduction

The Amazon Picking Challenge 2016 (APC)[1] required teams to solve picking and stowing tasks in a warehouse scenario. It provided a platform to compare state-of-the-art approaches to perception, motion planning, mechanics, and overall system design in the context of cluttered bin picking. In two separate tasks, a subset of 39 objects was either stowed from an unordered pile into an Amazon shelf, or retrieved from the shelf to fulfill an order.

A key challenge for perception systems in warehouse contexts are rapidly changing object sets. To be able to work with object categories varying daily, long training times have to be avoided. Also, the work needed to capture training examples needs to be kept minimal. For this reason, our approach uses extensively pretrained models, which are finetuned to the target domain on few examples. We find that this variant of transfer learning is well suited for this task.

The design of our APC system (see Fig. 1), which consists of a 6-DOF manipulator arm equipped with two RGB-D cameras and a flexible suction gripper, is described in Schwarz *et al.* [1].

II. Related Work

Aside from general bin picking works (e.g. [2]–[5]), there are reports from the Amazon Picking Challenge 2015 [6], [7]. Correll *et al.* [8] provides a nice summary and analysis of the approaches the teams used in 2015.

The idea of finetuning a pretrained network on few examples of the target domain is not new. For example, R-CNN [9] starts with a network trained for image classification on ImageNet and finetunes it for object detection. Pinheiro and Collobert [10] train a semantic segmentation network using large amounts of image-level labels. Schwarz *et al.* [11] use pretrained features and depth preprocessing for RGB-D object recognition and pose estimation.

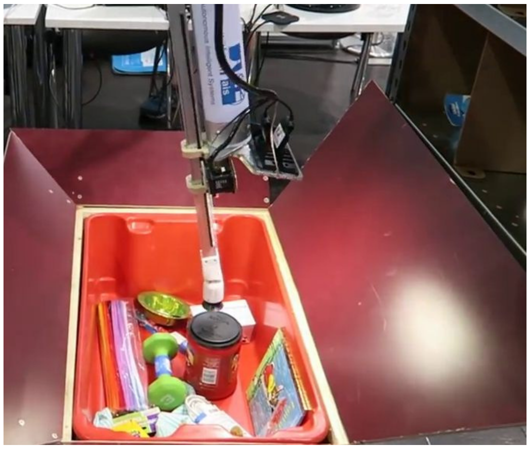

Fig. 1. Our system picking items from the tote at the APC 2016.

III. Object Perception

We adapted state-of-the-art object detection and semantic segmentation methods to our domain. The object detection approach outputs bounding boxes and object class probabilities, which is helpful in finding objects. Semantic segmentation is necessary to get a better understanding of the accessible object geometry in order to find a good grasp/suction spot.

For object detection, we extend the DenseCap network [12], which is originally designed for dense captioning, i.e. providing detailed textual descriptions of interesting regions in the input image. Similar to Faster R-CNN [13], an integrated region proposal network generates proposal boxes, which are then “focussed” using ROI pooling[2] and classified. Here, we make use of two stages of pretraining: The underlying CNN backbone network VGG-16 [14] was trained on ImageNet. The full DenseCap architecture was then finetuned on the Visual Genome dataset [15]. Finally, we adapt the network to output class probabilities and finetune again on our APC dataset.

Since depth measurements are available, we adapted the network to be able to make use of the additional modality (see Fig. 2). RGB and depth are processed independently up to the region proposal layer, were feature maps are concatenated. Due to the lack of large-scale annotated depth

University of Bonn, {schwarz|behnke}@ais.uni-bonn.de

[1] http://amazonpickingchallenge.org/

[2] More precisely, DenseCap uses ROI bilinear interpolation.

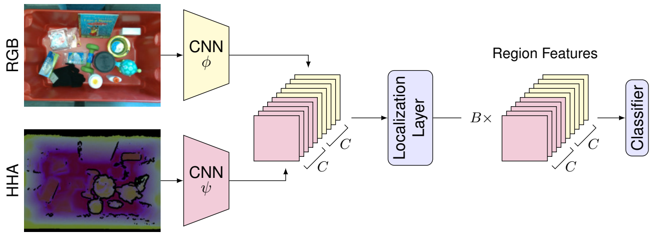

Fig. 2. Object Detection architecture for incorporating depth measurements. Pretrained CNNs ϕ and ψ are used for feature extraction. Each of the B object proposals are then classified using a classification head.

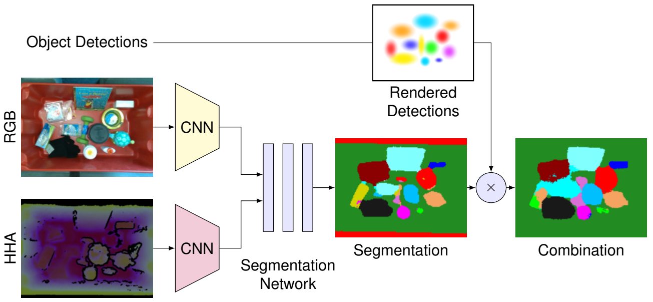

Fig. 3. Network architecture for semantic object segmentation and combination with object detection.

database which would be comparable to ImageNet for RGB images, there is no suitable pretrained CNN for depth feature extraction. Instead, we follow the Cross Modal Distillation approach [16], which trains a depth CNN to imitate features generated by a pretrained RGB CNN on unlabeled RGB-D sequences. In order to re-use the 3-channel VGG architecture, the popular HHA encoding [17] is used.

For estimating the object contour more precisely, we adapt our previous semantic segmentation network [18] to our application. Again, the underlying OverFeat [19] CNN was pretrained on ImageNet, and additional segmentation layers are then finetuned on our dataset. The full network architecture is illustrated in Fig. 3.

We also investigated a simple pixel-level combination of semantic segmentation and rendered gaussians from object detectionion which gave small but consistent gains in segmentation accuracy.

More details and thorough evaluation of our perception pipeline is available in Schwarz *et al.* [20].

IV. Results

A. Object Detection and Segmentation

We performed quantitative evaluations on the dataset we captured for training. It consists of 190 shelf frames, and 117 tote frames. An exemplary tote frame is shown in Fig. 4 along with the corresponding object detection and segmentation results. The dataset frames show the shelf and tote filled as they would during the competition, with manually created region labels. As far as we know, the number of frames is quite low in comparison to other teams. We attribute this achievement to the transfer learning approach.

TABLE I
Final object detection results on the APC dataset.

	mAP		F1
Dataset	Uninformed	Informed	
Shelf	0.878	0.912	0.798
Tote	0.870	0.887	0.779

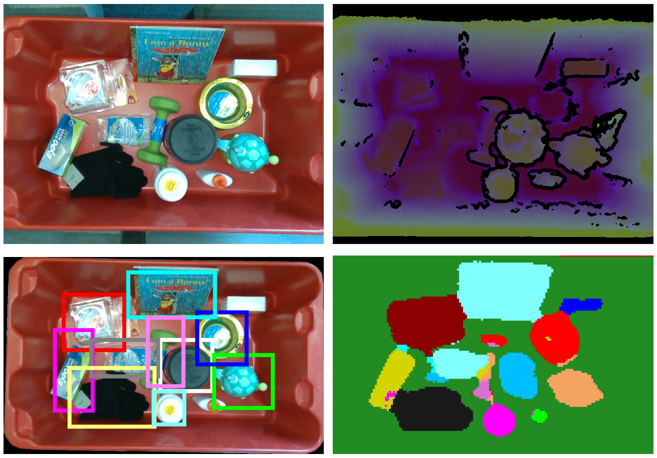

Fig. 4. Object perception example. Upper row: Input RGB and HHA depth frames. Lower row: Object detection and semantic segmentation results (colors are not correlated).

TABLE II
F1 scores for semantic segmentation.

	Shelf		Tote	
Method	Uninf.	Inf.	Uninf.	Inf.
Raw depth	0.713	0.735	-	-
HHA depth	0.780	0.813	0.817	0.839
Det+Seg[1]	0.795	0.827	0.831	0.853

[1] Object Detection + Segmentation.

For object detection, mAP and localization F1 scores (see [20]) are reported in Table I. Table II shows semantic segmentation scores. To migitate problems due to the low number of test examples, all scores are averaged over a five-fold cross validation split of the dataset.

B. Amazon Picking Challenge 2016

In addition to the good quantitative results on our dataset, our system was very successful at the APC 2016, resulting in a second place in the stow task and third place in the pick task. Videos of both runs are available[3]. While the pick task mostly suffered from dropped items, which resulted in incorrect output locations, the stow task was made difficult by a single misrecognition. This led the system to believe that a different object was in the tote, which could—of course—not be found. A backup strategy which would have attempted recognition of all known object classes failed because of an incorrect size threshold.

[3] Stow: `https://youtu.be/B6ny9ONfdx4`,
Pick: `https://youtu.be/q9YiD80vwDc`

V. Conclusion

The APC 2016 allowed our team to develop an efficient system for bin-picking in warehouse contexts. Applications of deep learning techniques in real-world robots are still rare, partly because of the large amounts of data required for training. Our stringent usage of pretrained models and unsupervised learning for feature transfer between modalities enabled us to use state-of-the-art deep learning methods. As a result, the system is data-efficient, learning from relatively few annotated examples. The system was proven at APC 2016 and evaluated in more detail on our APC dataset.

References

[1] M. Schwarz, A. Milan, C. Lenz, A. Munoz, A. S. Periyasamy, M. Schreiber, S. Schüller, and S. Behnke, "NimbRo Picking: Versatile part handling for warehouse automation," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2017.

[2] M. Nieuwenhuisen, D. Droeschel, D. Holz, J. Stückler, A. Berner, J. Li, R. Klein, and S. Behnke, "Mobile bin picking with an anthropomorphic service robot," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2013, pp. 2327–2334.

[3] C. Martinez, R. Boca, B. Zhang, H. Chen, and S. Nidamarthi, "Automated bin picking system for randomly located industrial parts," in *IEEE International Conference on Technologies for Practical Robot Applications (TePRA)*, 2015.

[4] K. N. Kaipa, A. S. Kankanhalli-Nagendra, N. B. Kumbla, S. Shriyam, S. S. Thevendria-Karthic, J. A. Marvel, and S. K. Gupta, "Addressing perception uncertainty induced failure modes in robotic bin-picking," *Robotics and Computer-Integrated Manufacturing*, vol. 42, pp. 17–38, 2016.

[5] K. Harada, W. Wan, T. Tsuji, K. Kikuchi, K. Nagata, and H. Onda, "Iterative visual recognition for learning based randomized bin-picking," *ArXiv:1608.00334*, 2016.

[6] C. Eppner, S. Höfer, R. Jonschkowski, R. Martín-Martín, A. Sieverling, V. Wall, and O. Brock, "Lessons from the Amazon Picking Challenge: Four aspects of building robotic systems," in *Robotics: Science and Systems (RSS)*, Jun. 2016.

[7] K.-T. Yu, N. Fazeli, N. Chavan-Dafle, O. Taylor, E. Donlon, G. D. Lankenau, and A. Rodriguez, "A summary of team MIT's approach to the Amazon Picking Challenge 2015," *ArXiv:1604.03639*, 2016.

[8] N. Correll, K. E. Bekris, D. Berenson, O. Brock, A. Causo, K. Hauser, K. Okada, A. Rodriguez, J. M. Romano, and P. R. Wurman, "Analysis and observations from the first amazon picking challenge," *IEEE Transactions on Automation Science and Engineering*, 2016.

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 580–587.

[10] P. O. Pinheiro and R. Collobert, "From image-level to pixel-level labeling with convolutional networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015.

[11] M. Schwarz, H. Schulz, and S. Behnke, "RGB-D object recognition and pose estimation based on pre-trained convolutional neural network features," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2015, pp. 1329–1335.

[12] J. Johnson, A. Karpathy, and L. Fei-Fei, "DenseCap: Fully convolutional localization networks for dense captioning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[13] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2015, pp. 91–99.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[15] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. Li, D. A. Shamma, M. S. Bernstein, and F. Li, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International Journal of Computer Vision (IJCV)*, vol. 123, no. 1, pp. 32–73, 2017.

[16] S. Gupta, J. Hoffman, and J. Malik, "Cross modal distillation for supervision transfer," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2827–2836.

[17] S. Gupta, R. Girshick, P. Arbelaez, and J. Malik, "Learning rich features from RGB-D images for object detection and segmentation," in *European Conference on Computer Vision (ECCV)*, 2014.

[18] F. Husain, H. Schulz, B. Dellen, C. Torras, and S. Behnke, "Combining semantic and geometric features for object class segmentation of indoor scenes," *IEEE Robotics and Automation Letters*, vol. 2, no. 1, pp. 49–55, May 2016, ISSN: 2377-3766.

[19] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *ArXiv:1312.6229*, 2013.

[20] M. Schwarz, A. Milan, A. S. Periyasamy, and S. Behnke, "RGB-D object detection and semantic segmentation for autonomous manipulation in clutter," *Accepted for International Journal of Robotics Research*, 2017.