Additional results for: Category-level recognition and pose estimation of small containers for human-to-robot handovers

Xavier Weber, Alessio Xompero, Andrea Cavallaro

I. INTRODUCTION

We show and discuss more detailed results of our models under comparison for 2D detection on the three subsets of images from the CCM dataset [1], and for 3D object detection and pose estimation on the test sets of the SOM dataset.

A. 2D object detection

We show the analysis of precision, recall, and Average Precision (AP) for each of the three main categories, namely *box*, *stem*, and *non-stem*, when varying the threshold, τ on the 2D intersection over union (IoU). We also compare our three retrained models, namely Ours-T, Ours-H, and Ours-HH, against Mask R-CNN [2] and the original NOCS [3]. Fig. 1 shows the results on *Tabletop (no hand)*, Fig. 2 on *Manipulating*, and Fig. 3 on *Handing over*.

Overall, precision, recall, and AP are at highest values up to $\tau = 75\%$ for our models and Mask R-CNN across the three object categories and the three test sets. NOCS has low AP performance, even at $\tau = 50\%$ for *stem* and *non-stem*, especially for handheld objects. Our models achieve higher AP than Mask R-CNN and NOCS, especially on *Tabletop (no hand)* and *Handing over* for *box* and *stem*, due to the higher recall. *Manipulating* is the more challenging set as there are hand occlusions and occlusions by other objects (i.e the pitcher), as shown in the lower performance by all models on this set. For *non-stem*, Mask R-CNN achieves similar recall ($> 75\%$) to NOCS-T, but worse recall ($< 60\%$) for the *stem* category, whereas higher precision is achieved for *non-stem*. On *Manipulating* and *Handing over*, Ours-H and Ours-HH outperform Mask R-CNN on all performance measures thanks to the training on synthetic hands, and therefore generalise better to real, handheld objects. Mask R-CNN achieves $< 20\%$ AP on *box* in all the three sets, especially because of the lower recall than precision. This is due to the fact that Mask R-CNN is not trained on actual images of boxes and we associated the category *book* with *box*.NOCS is not able to detect objects from the *box* category, as it has not trained on this or similar object categories. Ours-HH has higher precision for *box* for *Manipulating* and *Approaching* than Ours-H, as we observe that Ours-H often confuses the person and hand for a *box*. However, Ours-H has higher recall for *non-stem* for all test sets.

X. Weber, A. Xompero, and A. Cavallaro are with Centre for Intelligent Sensing, Queen Mary University of London, United Kingdom {x.weber,a.xompero,a.cavallaro}@qmul.ac.uk

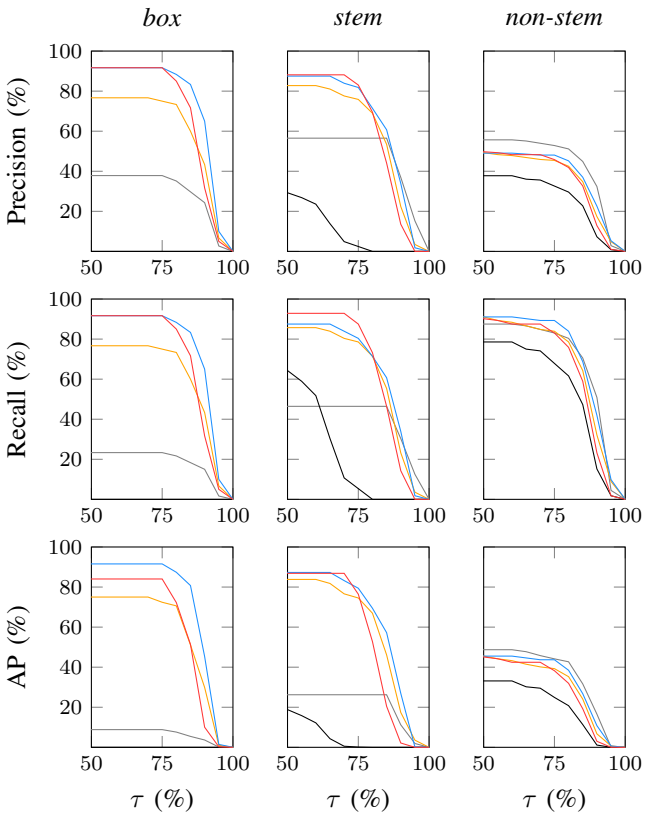

Fig. 1. Comparison of *Mask R-CNN [2]*, *NOCS [3]*, *Ours-T*, *Ours-H*, *Ours-HH* for 2D object detection results in terms of Precision, Recall, and Average Precision (AP), while varying the threshold on the Intersection over Union (τ) on CCM *Tabletop (no hand)*.

B. 3D object detection and pose estimation

We compare our re-trained models, as well as against the original NOCS, on the three test sets of SOM, namely *Tabletop (no hand)*, *Tabletop (handheld)*, and *Mid-air (handheld)*. To extract the object poses in real metric scale, our models use EPnP [4] and the mean point from the object's depth point cloud, while NOCS uses Umeyama's algorithm [5]. Note that NOCS was not trained on the *box* category and hence is not able to detect boxes. For *non-stem* and *stem*, we count a detection as correct when NOCS predicts the *mug* label. Moreover, we show in Fig. 4 the 48 synthetic objects collected from ShapeNetSem [6] and used in SOM. Note that images belonging to the objects in the last two columns are used to form the test sets.

For the *box* category (Fig. 5), only our models are compared and they share similar behaviour when varying

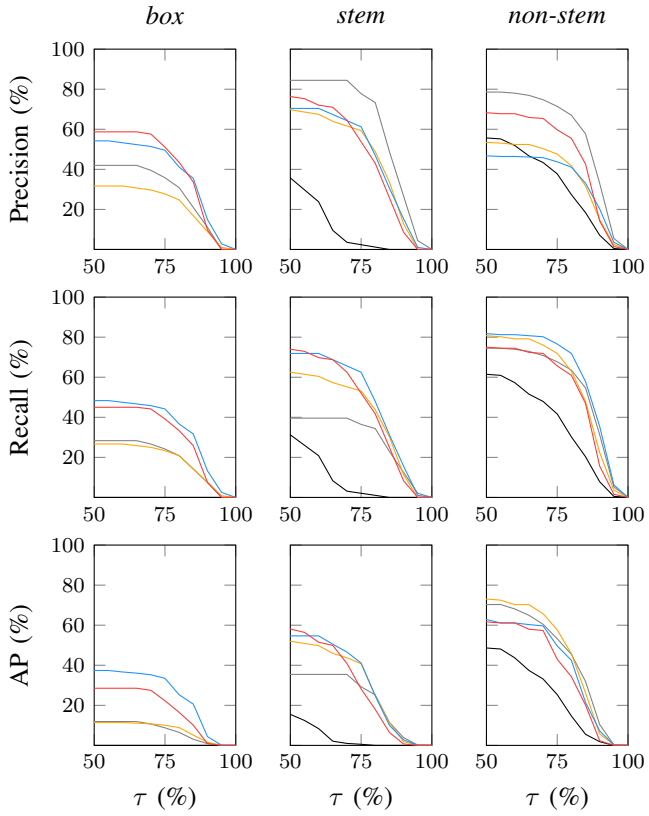

Fig. 2. Comparison of *Mask R-CNN [2]*, *NOCS [3]*, *Ours-T*, *Ours-H*, *Ours-HH* for 2D object detection results in terms of Precision, Recall, and Average Precision (AP), while varying the threshold on the Intersection over Union (τ) on CCM *Manipulating*.

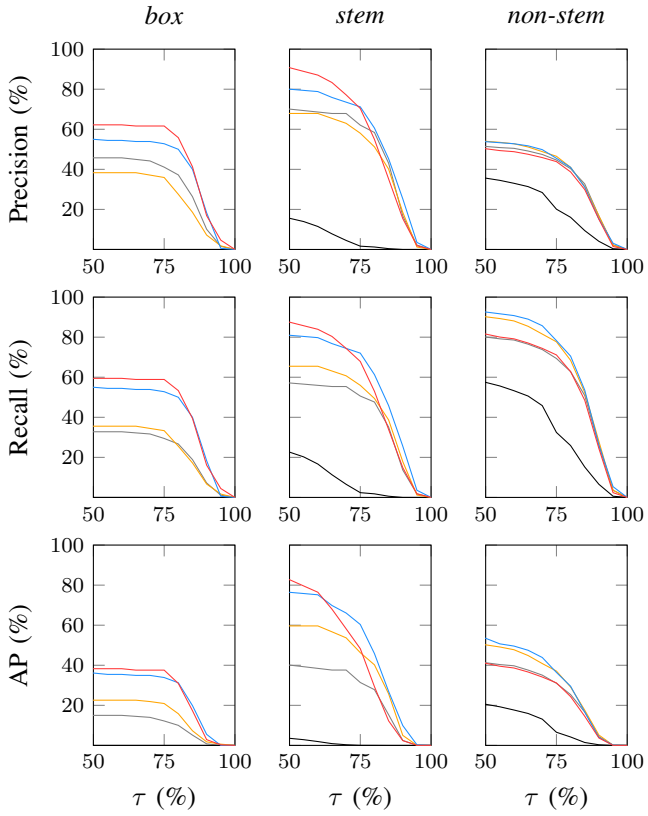

Fig. 3. Comparison of *Mask R-CNN [2]*, *NOCS [3]*, *Ours-T*, *Ours-H*, *Ours-HH* for 2D object detection results in terms of Precision, Recall, and Average Precision (AP), while varying the threshold on the Intersection over Union (τ) on CCM *Handing over*.

the thresholds on the IoU, and the translation and rotation components of the object pose. The AP already drops to less than 50% when the threshold $\eta \leq 50$, showing how challenging is the accurate detection of objects belonging to this class. By fixing $\eta = 10$ and $\psi = 360°$, AP decreases from 100% to almost 0% when limiting the translation error from 9 cm to 3 cm, on all test sets. By fixing then $\phi = 5$ cm and varying the threshold on the rotation component from 0 to 60°, we can observe how the AP is always lower than 40%, decreasing to 0 at about 10°, also on all test sets. Compared to Ours-H/HH, Ours-T achieves a lower AP for 3D object detection at low thresholds ($\eta \leq 30$), while being higher for $\eta > 30$. Overall, the lower performance compared to other categories is likely due to the presence of a box with similar width and depth dimensions, therefore decreasing the accurate predictions of the 3D normalised coordinates.

For the *non-stem* category (Fig. 6), the AP of the models decreases when the threshold on the 3D IoU is higher than 70%, while NOCS has lower AP (about 80%) than our models when $\eta \leq 70$. Our models also outperform NOCS in the pose estimation, even considering our models detect correctly many more objects, and show high AP (about 90%) up to $\psi = 10°$ and $\phi = 5$ cm, for objects without occlusion. On the other two test sets with hand occlusions, all models decrease in correctly detecting the objects in 3D. Importantly, Ours-H/HH handle better hand occlusions than Ours-T and NOCS, achieving an AP higher than 75% when $\eta \leq 70\%$, due to training on synthetic hands. A decrease in AP when the objects are held in the mid-air is present for all models, because these objects are challenging due to occlusion and a greater variation in rotation.

For the *stem* category (Fig. **??**), NOCS is the worse in all the three subsets, achieving an AP lower than 20% for 3D object detection. This is caused by the fact that NOCS can detect the container but often ignores the stem of this object, only predicting the upper part of the container, which is more similar to the *non-stem* category. Relative to the correct 3D IoU at $\eta = 10\%$, also the AP of the pose estimations is worse than our models and lower than 10% when the maximum rotation error is 15° and maximum translation error fixed at 5 cm. Ours-H and Ours-HH outperform Ours-T in all test sets when evaluating 3D object detection, showing that training the models with hand occlusion also improves the recognition of parts like a stem. In this case, the benefit is especially visible for Ours-HH that achieves about 80 AP, 20 AP higher than Ours-H, in both *Tabletop (handheld)* and *Mid-air (handheld)*, when the threshold on the 3D IoU is $< 70\%$.

For all object categories, the performance substantially deteriorates for all models when introducing the hand occlu-

Fig. 4. Objects of our SOM dataset: *boxes* (top), *stem* (middle), *non-stem* (bottom).

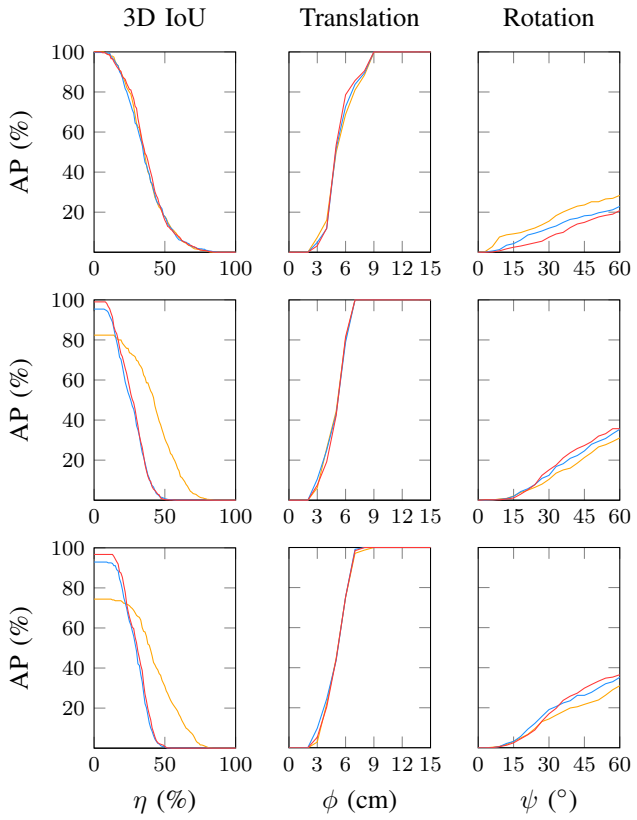

Fig. 5. Comparison of Average Precision (AP) results between *Ours-T* (—), *Ours-H* (—), *Ours-HH* (—) for *box* on *SOM - Tabletop (no hand)* (top), *SOM - Tabletop (handheld)* (middle), and *SOM - Mid-air (handheld)* (bottom). We measure Average Precision (AP) for 3D object detection while varying the 3D IoU threshold, η. We evaluate pose estimation while varying ϕ for translation (with maximum error in rotation fixed at 360°), and ψ for rotation (with maximum error in translation fixed at 5 cm), while the 3D IoU is fixed at 10%.

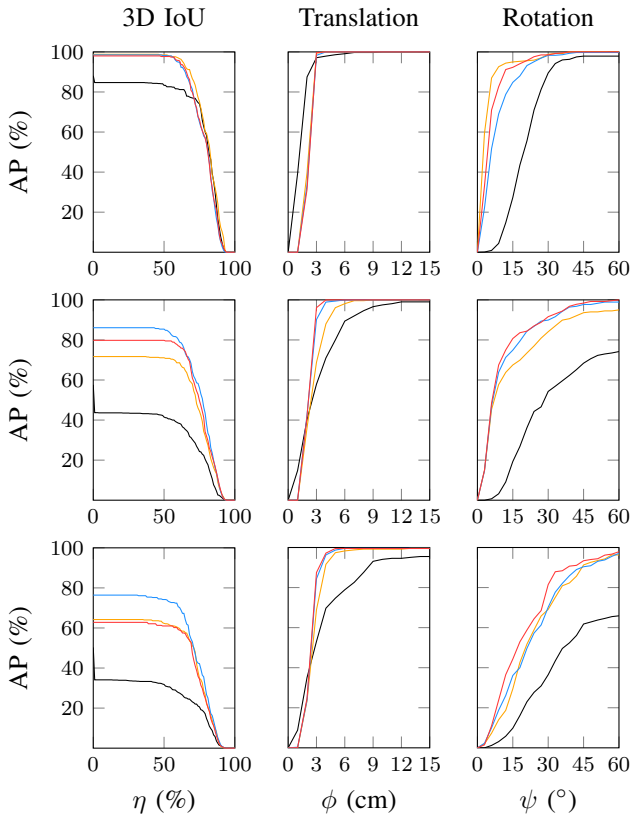

Fig. 6. Comparison of Average Precision (AP) results between *NOCS [3]* (—), *Ours-T* (—), *Ours-H* (—), *Ours-HH* (—) for *non-stem* on *SOM - Tabletop (no hand)* (top), *SOM - Tabletop (handheld)* (middle), and *SOM - Mid-air (handheld)* (bottom). We measure Average Precision (AP) for 3D object detection while varying the 3D IoU threshold, η. We evaluate pose estimation while varying ϕ for translation (with maximum error in rotation fixed at 360°), and ψ for rotation (with maximum error in translation fixed at 5 cm), while the 3D IoU is fixed at 10%.

sions and reducing the maximum allowed rotation error, and further reduces when the object is rotated more in *Mid-air*. Specifically for NOCS, the model is not able to accurately detect and segment the objects due to the removal of the deep layers in the pre-trained Mask R-CNN, where the representations for detecting humans are more likely to be learned, and also not fine-tuning on the human category or on images with hands during training. Due to training on synthetic hands, Ours-H and Ours-HH improve the object segmentation, reduce the false positives, and hence improve the 3D object detection.

ACKNOWLEDGMENT

This work is supported by the CHIST-ERA program through the project CORSMAL, under UK EPSRC grant EP/S031715/1.

REFERENCES

[1] A. Xompero, R. Sanchez-Matilla, R. Mazzon, and A. Cavallaro, "CORSMAL Containers Manipulation," (1.0) [Data set]. Queen Mary University of London. [Online]. Available: http://corsmal.eecs.qmul.ac.uk/containers_manip.html

[2] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017.

[3] H. Wang, S. Sridhar, J. Huang, J. Valentin, S. Song, and L. J. Guibas, "Normalized object coordinate space for category-level 6d object pose and size estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019.

[4] V. Lepetit, F. Moreno-Noguer, and P. Fua, "EPnP: An accurate O(n) solution to the PnP problem," *Int. J. Comput. Vis.*, vol. 81, no. 2, pp. 155–166, 2009.

[5] S. Umeyama, "Least-squares estimation of transformation parameters between two point patterns," *IEEE Computer Architecture Letters*, vol. 13, no. 04, pp. 376–380, 1991.

[6] M. Savva, A. X. Chang, and P. Hanrahan, "Semantically-enriched 3D models for common-sense knowledge," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2015.