https://doi.org/10.31256/Sp7Gn7W

Exploring Rotated Object Detection Models for Antipodal Robotic Grasping

Valerija Holomjova
University of Aberdeen
Aberdeen, Scotland UK
v.holomjova.21@abdn.ac.uk

Pascal Meißner
University of Aberdeen
Aberdeen, Scotland UK
pascal.meissner@abdn.ac.uk

***Abstract*—Current deep learning approaches used by robotic grasping systems for predicting multiple valid grasps across various objects from images have achieved great results, but often stem from object detectors that were originally designed for predicting horizontal bounding boxes. Since 2D grasp poses are more naturally represented by oriented bounding boxes, in this paper, we explore the suitability of three top-performing rotated object detectors as they are composed of modules tailored for encoding rotated object features more precisely. The performance of the oriented detectors is compared against an effective grasp detection model architecture from literature on two publicly available grasping datasets. Results show that oriented detectors obtained comparable grasp accuracy scores on both datasets, whilst being more capable of producing confident and diverse sets of grasps. Code is available at *https://github.com/valerija-h/exploring_rotated_object_detection_models*.**

***Index Terms*—Grasping, Deep Learning for Robotics, Perception for Grasping, Computer Vision for Automation**

I. Introduction

Over the years, deep learning models have enabled antipodal robotic grasping systems to infer stable grasps across various objects in their environment, permitting them to perform assistive or industrial tasks such as binning, sorting and assembling. Techniques that allow such systems to perform rapid yet stable grasps on a diverse set of objects without an object model in unstructured environments remain an active area in research.

Similar to the notation of a bounding box $b = (x, y, w, h)$, a 2D grasp pose can be represented as a grasp rectangle $g = (x, y, w, h, \theta)$, where (x, y) is the centre point of the gripper, w, h is the gripper opening and size respectively and θ denotes the orientation of the gripper w.r.t. to the horizontal axis (Fig. Ia). Using such notation, one effective approach from literature for training a multi-object grasp detection model is by using a standard object detector (e.g. Faster R-CNN [1]), and replacing the object classes it predicts with an orientation class r [2], [3]. Due to the symmetry of the grippers, values of orientation θ lie in the range [-$\frac{\pi}{2}$, $\frac{\pi}{2}$] and can be discretized into N classes s.t. the set of possible orientation classes is $R = \{r_1, ..., r_N\}$ with an additional class c denoting an invalid grasp.

More recently, researchers in Computer Vision have been challenged with the task of rotated object detection involving precisely detecting vehicles from aerial images resulting in a rise of detector models tailored for predicting oriented bounding boxes. The S^{2}A-Net [4] is a single-shot alignment network that aligns deep convolutional features to rotated anchors. ReDet [5] has the ability to encode both rotation equivariant and invariant features. The Oriented R-CNN [6] is a two-stage oriented detector that uses an oriented Region Proposal Network (RPN) to generate quality oriented proposals.

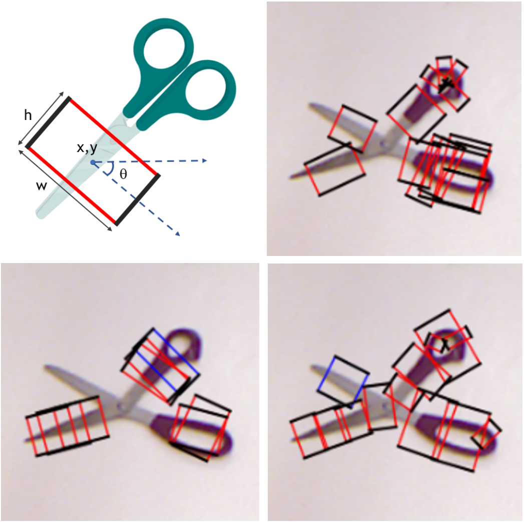

Fig. 1. (a) **Top left** illustrates an example of a grasp rectangle on an object. (b) **Top right** shows a sample RGB image from the Cornell dataset [7] with hand-annotated grasp rectangles. (c) **Bottom left** shows predictions from the baseline model. (d) **Bottom right** shows the predictions from the Oriented R-CNN [6] model. Grasp predictions with confidence scores > 0.3 are shown where the blue grasps have the highest confidence.

Since standard object detectors used for grasp detection were originally intended for predicting horizontal bounding boxes, this paper seeks to explore whether recent oriented detectors would yield more precise grasps as they are tailored to encode rotated object features more accurately. Our **key contribution** is to compare the performance of three oriented detectors to a baseline grasp detection model for predicting suitable grasp rectangles on objects from input RG-D images. The baseline model will be built from a Faster R-CNN using common techniques found in grasp detection literature. Two publicly available datasets (Cornell [7] and OCID [2]) containing depth and RGB images of diverse objects annotated

This research is funded by a studentship awarded by the School of Engineering at the University of Aberdeen, Scotland UK.

with multiple grasps will be used for training and evaluation.

II. Methodology

Each dataset was pre-processed into its required format and split image-wise into 80% training and 20% testing samples, where 10% of the training samples were used for validation and a seed value was set to ensure reproducibility. Images in the Cornell and OCID dataset were cropped to a size of $315 \times$ 315px and 590×460px respectively to remove unnecessary background noise or objects.

The baseline model was built and trained in PyTorch using a pre-trained Faster R-CNN, whereas the rotated object models were implemented using MMRotate [8], an open-source toolbox based on PyTorch. Each model was trained on RG-D images for a maximum of 5 epochs using their respective losses on an NVIDIA GeForce RTX 3070 with CUDA 11.3. An Adam optimizer with an initial learning rate of 0.0001 was used for training as well as a training batch size of 2. Random horizontal and vertical flips with a trigger rate of 25% were also used in the training pipelines.

III. Evaluation

Similar to previous literature [2], [3], the rectangle metric is used to calculate the grasp accuracy of each model on both datasets. The metric classifies a predicted grasp rectangle g_p as valid when evaluated against a ground truth grasp rectangle g_{gt} if both of the conditions below are met;

- The angle difference between g_p and g_{gt} is within $30°$.
- The Intersection over Union (IoU) score between g_p and g_{gt} is greater than 25%:

$$\text{IoU}(g_p, g_{gt}) = \frac{|g_p \cap g_{gt}|}{|g_p \cup g_{gt}|} > 0.25 \tag{1}$$

The predicted grasp with the highest confidence score is chosen for calculating the grasp accuracy.

IV. Results

Tables I and II report the grasp accuracy and average inference speed of each model on the Cornell and OCID datasets respectively. Apart from S^2A-Net, both tables show that all models achieved comparable grasp accuracy scores. The baseline model had the highest inference speed in the Cornell dataset which then dropped much more than the other models in the OCID dataset. This could be attributed to the fact that the OCID dataset has multi-object scenes whereas the Cornell dataset has single object scenes, which could suggest that oriented detectors are more efficient in multi-object scenes.

As depicted in Fig. Ic-d, qualitative results from the Cornell dataset show that the baseline model often identifies less diverse grasps on objects. Its predicted grasps also have lower confidence scores with riskier stability as the bounds of the grasp rectangle would sometimes collide with the object itself. In fact, the average confidence score of the baseline's valid grasps in the Cornell dataset is 38%, whereas the rotated object detectors have an average confidence score above 90% in both datasets. However, in the OCID dataset, the average confidence score of the baseline's valid grasps increases to 84%. Despite this, when inspecting predicted grasps of each model that obtained a confidence score above 30% on both datasets, the baseline model would very often not provide any suitable grasps on certain objects in both single-object and multiple-object settings making it difficult to determine a confidence threshold to use such that only quality grasps are considered.

TABLE I

Comparison of model performance on the Cornell dataset.

Model	**Grasp Accuracy** (%)	**Speed** (FPS)
Baseline	96.61	**18.1**
Oriented R-CNN [6]	**97.18**	13.1
ReDet [5]	94.35	10.4
S^2A-Net [4]	92.09	13.2

TABLE II

Comparison of model performance on the OCID dataset.

Model	**Grasp Accuracy** (%)	**Speed** (FPS)
Baseline	97.73	13.9
Oriented R-CNN [6]	97.17	14.3
ReDet [5]	**98.58**	11.5
S^2A-Net [4]	91.50	**14.4**

V. Conclusion

This paper explores the suitability of rotated object detectors for detecting 2D grasp poses from RG-D images. The models are compared to a baseline grasp detection model based on a traditional object detector model on two publicly available datasets. Results show that all models achieve comparable grasp accuracy on both datasets, but rotated object detectors appear to provide more confident and diverse stable grasps, especially in the presence of multi-object scenes. Future work involves evaluating these models in real-world experiments centred around grasping unseen objects using a robotic arm.

References

[1] S. Ren *et al.*, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in *Advances in Neural Information Processing Systems*, C. Cortes *et al.*, Eds., vol. 28, Curran Associates, Inc., 2015.

[2] S. Ainetter and F. Fraundorfer, "End-to-end Trainable Deep Neural Network for Robotic Grasp Detection and Semantic Segmentation from RGB," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2021, pp. 13 452–13 458.

[3] F. J. Chu, R. Xu, and P. A. Vela, "Real-world Multiobject, Multigrasp Detection," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3355–3362, 2018.

[4] J. Han *et al.*, "Align Deep Features for Oriented Object Detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2021.

[5] J. Han *et al.*, "ReDet: A Rotation-equivariant Detector for Aerial Object Detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 2786–2795.

[6] X. Xie *et al.*, "Oriented R-CNN for Object Detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 3520–3529.

[7] I. Lenz, H. Lee, and A. Saxena, "Deep Learning for Detecting Robotic Grasps," *The International Journal of Robotics Research (IJRR)*, vol. 34, no. 4-5, pp. 705–724, 2015.

[8] Y. Zhou *et al.*, "MMRotate: A Rotated Object Detection Benchmark using PyTorch," *arXiv preprint arXiv:2204.13317*, 2022.