Visual Annotation of Gripping Points for Robotic Manipulation

Caroline HULOT

May 27, 2025

Abstract

This project focuses on the annotation of gripping points in 2D images to create a high-quality dataset for robotic manipulation tasks. Initially, a small set of images was manually annotated using *Label Studio*, marking two gripping points per object. To accelerate the process, *YOLO* was used to generate additional annotations on new images. The automatically annotated data was then reviewed and refined to ensure accuracy. This approach balances precision and efficiency, enabling the rapid creation of an extensive annotated dataset.

Introduction

Image annotation is the process of labeling visual data to enable machines to interpret and interact with their environment. In robotics, this technique is foundational for training systems to recognize objects, locate gripping points, and perform precise manipulation tasks—ranging from industrial assembly and packaging to household automation and medical procedures. High-quality annotated datasets are essential for developing robust predictive models, bridging the gap between theoretical research and real-world applications.

A critical challenge in robotic grasping is the accurate identification of gripping points, especially in 2D images lacking depth information. Unlike 3D sensors, which provide spatial data, 2D images require sophisticated annotation strategies to compensate for the absence of geometric context. This limitation becomes pronounced when handling objects with diverse shapes, textures, or reflectivity. For instance, irregular surfaces or occlusions can hinder consistent annotation, directly impacting a robot's ability to generalize across unfamiliar objects.

Precise annotation enhances robotic efficiency across domains:

- **Manufacturing:** Annotated datasets enable defect detection and optimized material handling, improving assembly lines and packaging processes.[1]
- **Healthcare:** Robots utilizing annotated medical images assist in diagnostics and surgeries. For example, scientists have developed a small robot designed to detect and treat bowel cancer by navigating the digestive system and making 3D scans.[2]
- **Domestic Robotics:** Assistants rely on annotated gripping points to manipulate household objects autonomously, enhancing human-robot interaction.

By applying these technologies, we aim to create tangible solutions that enhance the interaction between robots and the physical world, leading to more efficient and effective robotic systems across various applications.

1 State of the Art

The field of robotic grasping has undergone significant transformation through advances in image annotation techniques, enabling machines to better interpret and manipulate objects in their environment. While early systems relied on geometric heuristics, modern approaches leverage deep learning to handle complex real-world scenarios. This evolution reflects the growing need for efficient, scalable solutions that maintain high precision.

1.1 Methods for Generating Annotated Data

Contemporary annotation employ three complementary approaches, each addressing specific needs in the data preparation process:

- **Manual annotation**: Manual annotation remains the gold standard for precision, with tools like *Label Studio* enabling human to meticulously label key features. However, the labor-intensive nature of this approach has driven the development of augmentation techniques (rotation, lighting adjustments) to maximize dataset utility from limited annotations [3].

- **Semi-automated methods**: Semi-automated methods mark a significant change in how annotations are created, combining human expertise with machine efficiency. [4]. This hybrid approach significantly accelerates annotation workflows without compromising accuracy.

- **Data synthesis**: Data synthesis through simulation environments offers another innovative solution.Systems like LabelFusion[5] create high-quality synthetic datasets with accurate annotations, enabling "sim-to-real"[6] transfer learning. These techniques are particularly valuable for scenarios where collecting real-world data would be impractical or dangerous.

1.2 Advanced 2D Annotation Techniques

The transition from geometric methods to data-driven approaches has revolutionized 2D grasp detection:

- Traditional geometric approaches, relying on contour detection and symmetry analysis [7], provided initial solutions but proved inadequate for complex objects, limiting their real-world applicability.

- Modern neural network architectures have overcome these limitations through learned feature representation. Object detection models like YOLO [8] and Faster R-CNN [9] provide real-time identification of potential grasp regions.These approaches excel because they:

  - Automatically learn relevant features from data rather than relying on hand-crafted rules.
  - Adapt to diverse object types through training on varied datasets.

1.3 Current Challenges and Emerging Solutions

The impact of these techniques is amplified by foundational datasets like ImageNet and CIFAR-10, which have enabled training of increasingly sophisticated models. Advanced applications like the "Show and Tell" image captioning system [10] demonstrate the remarkable capabilities of modern annotation systems.

Despite these advancements, significant obstacles remain in developing robust annotation systems:

- The fundamental limitation of 2D data persists, with the lack of depth information requiring complex compensation heuristics [11].

Innovative solutions are emerging to address these limitations:

- Multi-modal approaches combining RGB with depth estimation [12].

- Hybrid systems like the two-stage method [13] that combine optimization algorithms with neural networks.

The field continues to evolve rapidly, with each advancement building on previous work. From manual annotation to sophisticated neural networks, progress in image annotation methods has been instrumental in advancing robotic grasping capabilities.

2 Description of the Work

This project follows a systematic methodology to address the key challenges in annotating gripping points on 2D images. The goal is to create a dataset of images with annotations for gripping points.

2.1 Initial Manual Annotation with Label Studio

To establish a reference dataset, I began by manually annotating a set of simple-shaped objects using *Label Studio*.

Procedure:

- For each object, two annotation points (corresponding to optimal grasping points for a two-finger gripper) were marked.

- These points were placed according to geometric criteria (e.g., symmetry centers, stable edges) and physical constraints.

- **Example**: For a rectangle, points were positioned at the center of opposite sides.

Figure 1: Gripping point on a rectangle.

Manual annotation ensures high initial precision, which is essential for training an automatic model.

2.2 Automatic Annotation with YOLO

After creating the initial annotated dataset, i used *YOLO* (You Only Look Once) to generalize annotations to more varied objects.

- **Implementation**:
  - **Pre-trained model**: YOLOv8 (`https://github.com/ultralytics/ultralytics`), selected for its speed and compatibility with real-time detection tasks.
  - **Training command (Linux)**: `python3 train.py -data dataset.yaml -weights yolov8s.pt -epochs 50 -imgsz 640`

    The model was trained for 50 epochs, meaning the entire dataset was processed 50 times. This number was chosen to ensure sufficient learning while avoiding overfitting, balancing accuracy and training efficiency.
  - **Process**:
    1. The model was trained on manually annotated images
    2. It then predicted grasping points on new images by generating bounding boxes around candidate areas

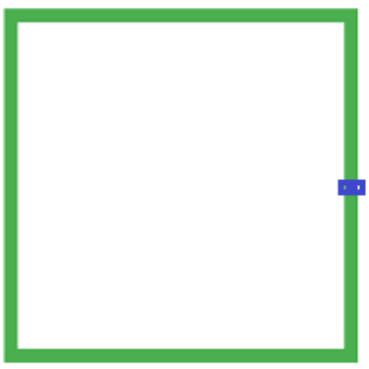

Figure 2: Automatic Annotation.

Figure 3: Automatic Annotation.

2.3 Validation and Adjustments

To ensure quality of automatic annotations:

- **Manual verification**: YOLO-annotated images were reviewed by myself
- **Adjustments**: False positives (e.g., points on unstable areas) were corrected in **Label Studio** and fed back into YOLO for improvement

3 Conclusion

Through this project, I manually annotated images and trained an AI model on these annotations to generate additional labels for new objects By combining manual annotation with *Label Studio* and automated annotation using *YOLO*, I achieved a balance between precision and efficiency, significantly accelerating the dataset creation process. The integration of a validation step ensured that automatically generated annotations remained accurate and reliable for further use.

However, some challenges remain. Using only 2D images makes it harder to understand the exact positions of gripping points. This could be improved by incorporating depth estimation techniques or multi-modal data. Additionally, the automatic annotation process showed limitations when dealing with complex object shapes or low-quality images . These factors often led to incorrect or imprecise gripping point predictions.

Future work could focus on expanding the dataset with a wider variety of object shapes and challenging image conditions, such as poor lighting, occlusions, different resolutions, and varying object orientations. Improving these aspects will make gripping point detection more accurate and help develop robotic grasping systems that are both efficient and precise, Reducing the gap between computer vision and real-world object handling

3.1 My work

Throughout this project, I have:

- Annotated images and objects for gripping point detection using *Label Studio* .
- Successfully transitioned from manual to automated annotation while maintaining quality control standards.
- Used *YOLO* to automate the annotation of gripping points.
- Adapted standard object detection techniques (YOLO) to the specific challenge of gripping point identification.
- Established annotation protocols that consider real-world robotic constraints, particularly for two-finger grippers.

References

[1] S. Dikici and R. John Robinson, "Automated defect detection using image recognition in manufacturing," *Journal of Data Science and Intelligent Systems*, 2024.

[2] N. e. a. Greenidge, "A mussel-inspired magnetic soft robot for minimally invasive procedures," *Science Robotics*, 2025.

[3] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," *arXiv preprint*, vol. arXiv:1712.04621, 2017.

[4] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, and L. Gustafson, "Segment anything," *arXiv preprint*, vol. arXiv:2304.02643, 2023.

[5] P. Marion, P. Florence, L. Manuelli, and R. Tedrake, "Label fusion: A pipeline for generating ground truth labels for rgb-d dataset," *arXiv preprint*, vol. arXiv:1707.02732, 2017.

[6] S. James, A. J. Davison, and E. Johns, "Transferring sim-to-real in robotics with deep reinforcement learning," *IEEE Transactions on Robotics*, vol. 38, no. 5, pp. 2752–2765, 2022.

[7] X. Wang, J. Li, and Y. Zhang, "Memory efficient grasping point detection of nontrivial objects," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1555–1562, 2020.

[8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *arXiv preprint*, vol. arXiv:1506.02640, 2016.

[9] 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, pp. 1137–1149, 2015.

[10] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3156–3164, 2015.

[11] D. Alvarez, G. Barros, F. Garcia, F. Della Santina, and Y. Chen, "Affordance-based grasping point detection using graph convolutional networks," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 873–880, 2021.

[12] J. Park, S. Kim, and T. Lee, "Depth-informed grasp prediction for robotic manipulation," *IEEE Transactions on Robotics*, vol. 39, no. 5, pp. 1256–1268, 2023.

[13] H. Chu, X. Zhang, and J. Lee, "A hybrid approach for improved grasp detection," *Robotics and Automation Letters*, vol. 9, no. 3, pp. 2104–2110, 2024.