Dexterous Functional Manipulation for Articulated Objects

Junho Kim[1], Claire Chen[2], Jeannette Bohg[2]
[1] Rice University, [2] Stanford University

Background

Grasping with a **robotic hand** is a long-standing challenge. It can be categorized by factors such as object complexity, shape, and functionality. Our focus is on articulated objects, which have the following characteristics:

- Multiple affordances
- Revolute or prismatic joints
- Graspable
- Functional

Previous work has utilized various grippers to grasp and manipulate objects with differing complexities in affordances and articulation. However, these studies often overlook the **functionality** of the object, failing to demonstrate its proper use. To the best of our knowledge, our work is the first to utilize multi-affordance mask alignment for evaluation, complementing the robustness of stability and functionality with Intersection over Union (IoU) and Chamfer Distance metrics. Our method effectively assesses grasp functionality and execution using policy-based learning, enabling robotic hands to grasp and functionally use objects.

Experimental Design

We decompose dexterous manipulation into **two stages**: the **grasping** stage and the **post-grasp** stage, which involves using the articulated object.

Fig. 1: Stage 1.1. We use the ZED stereo camera for real-world object segmentation. Using camera calibration, we calculate depth from intrinsic and extrinsic parameters to generate a partial point cloud. A stereo camera setup simplifies the pipeline but sacrifices full object mesh reconstruction.

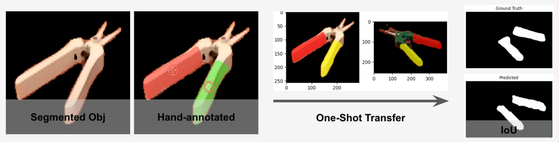

Fig. 2: Stage 1.2. The 2D segmented objects are hand-annotated to establish reference ground truth and object ground truth. The reference ground truth serves as the one-shot transfer target input, while the object ground truth is used to evaluate the Intersection over Union (IoU) of the affordance transfer method [2].

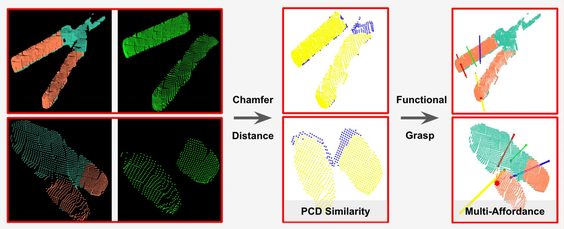

Fig. 3: Stage 1.3. The partial point cloud of objects is overlapped with ground truth and predicted affordance mask regions. The similarity between these masks in 3D space is evaluated using Chamfer Distance. Feasible functional grasps, generated by the optimization-based method [1] using object geometry, are filtered using a multi-affordance metric.

Experimental Design (Continued)

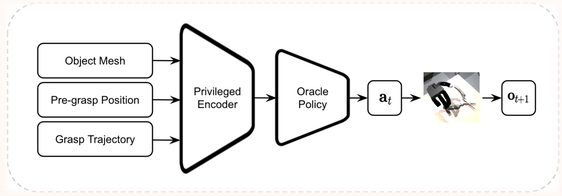

Fig. 4: Stage 2. An overview of our approach. We train an oracle policy in simulation using reinforcement learning. We jointly optimize the privileged encoder using Proximal Policy Optimization (PPO).

Methodology

Evaluation Metrics

Intersection of Union	$J(A,B) = \frac{\lvert A \cap B \rvert}{\lvert A \cup B \rvert}$
Chamfer Distance	$\text{chamfer}(P_1, P_2) = \frac{1}{2n}\sum_{i=1}^{n}\lvert x_i - \text{NN}(x_i, P_2)\rvert + \frac{1}{2m}\sum_{j=1}^{n}\lvert x_j - \text{NN}(x_j, P_1)\rvert$
Multi-Affordance Contact Alignment	$\text{MACA} = \frac{1}{N}\sum_{i=1}^{N}\delta(f_i, R_i)$

Reinforcement Learning Framework

Proximal Policy Optimization	$L(\theta) = \mathbb{E}_t\left[\min\left(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\right)\right]$
Policy Network	$r_t(\theta) = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)}$ ⟵ $\pi_\theta(a_t \mid s_t) = \text{Neural Network}_\theta(s_t)$
Network State Input	$s_t = [\text{fingertip positions, joint angles, object position, contact forces, object state}]$
Reward Function	$R(s_t, a_t) = w_1 \cdot \text{Grip Success} + w_2 \cdot \text{Squeeze Force} - w_3 \cdot \text{Slip Penalty}$

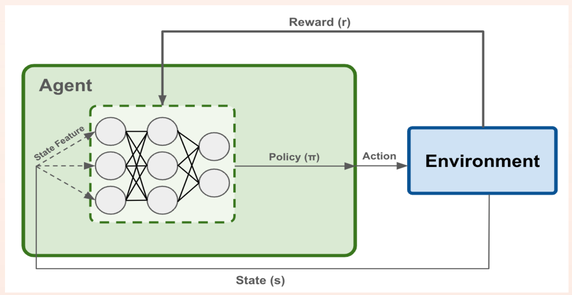

Fig. 5: The agent learns by interacting with the environment: it takes an action based on the current state, receives a reward, and updates its policy to improve future decisions.

- We plan to use PPO to improve the policy in a stable and effective way. This method carefully adjusts the policy by considering both the potential benefits and limitations of each action.
- The reward function is designed to optimize grasping performance by rewarding grip success and squeeze force while penalizing slippage.
- Leveraging Isaac Gym's high-fidelity simulation environment, PPO efficiently trains the policy in parallel across multiple environments, accelerating the learning process for complex manipulation tasks.

Results

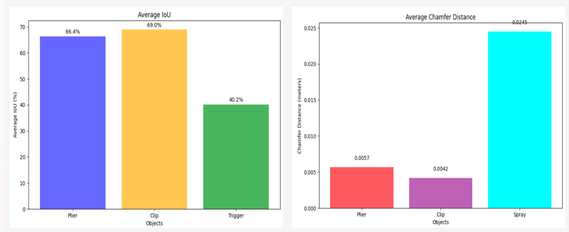

Fig. 6: The figure presents a comparison of model performance using two metrics: Intersection over Union (IoU) on the left and Chamfer Distance on the right.

- 2D-based affordance generalization is effective within the same class of objects, accurately capturing their semantics.
- Chamfer distance is used to evaluate the similarity between partial point clouds of ground truth and predicted affordance masks.
- IoU (Intersection over Union) and Chamfer distance are complementary in their relationship:
  - Objects with clear semantics (e.g., pliers, clips) exhibit higher IoU and lower Chamfer distance.
  - Objects in the spray category (e.g., spray, drill, glue-gun) show the opposite trends, with lower IoU and higher Chamfer distance, due to the difficulty in generalizing their affordance regions in 2D space.

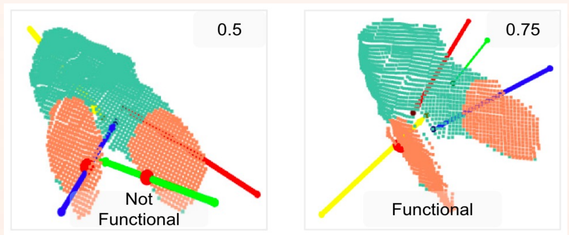

Fig. 7: A side-by-side comparison of two grasps: a stable, non-functional grasp (left), a stable, functional grasp (right). The vectors yellow, red, green, and blue represent the thumb, index, middle, and ring fingers, respectively.

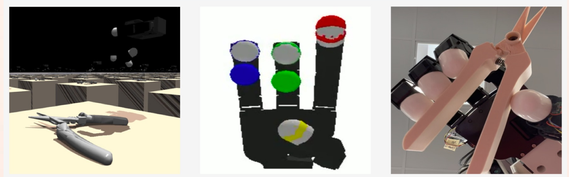

Fig. 8: (a) The current Isaac Gym environment featuring a plier and the Allegro hand. (b) Color-coded tracking of Allegro hand fingertips, corresponding to vectors shown in Figure 7. (c) Real-world object manipulation using the Allegro hand and SpringGrasp.

- Affordance generalization and similarity between 3D affordance regions show reasonable accuracy compared to prior work [2].
- A multi-affordance alignment metric was developed to validate that stable grasps generated using the rigid body grasping method from [1] align with affordance regions.
- The metric distinguishes between stable, non-functional grasps and stable, functional grasps.
- This evaluation is critical for assessing grasp functionality beyond the common 'grasp and use' success rate.
- The novel evaluation criterion can also be integrated as a loss function to differentiate stable from functional grasps.

Conclusion

- Robotic manipulation often uses full object meshes or multi-view point clouds.
- Partial point clouds simplify the experimental setup but require a robust grasp generation method.
- Grasping is a significant part of solving functional grasping. Understanding object semantics and point level affordance is necessary.
- Developing a policy for robots to grasp small objects with precise, point-level affordance is challenging.
- Sophisticated hand movements are difficult to control through teleoperation, making the task non-trivial for imitation learning [3].
- Generalization to other tasks or unseen objects remains an unsolved challenge.

Future Works

- Integrate the multi-affordance alignment metric as a loss function within SpringGrasp.
- Perform quantitative and qualitative comparisons between functional grasping and robust grasping.
- Simulate grasp vectors in the Isaac Gym environment and integrate the grasping pipeline with Isaac Gym.
- Develop a reward function and train the policy.
- Conduct extensive testing within the simulation on various objects (e.g., pliers, clips, spray bottles, drills, etc.).
- 3D print objects from the simulation for real-world testing, ensuring that the scale from object to hand to the real world is accurate.
- Implement Sim2Real transfer for real-world application.

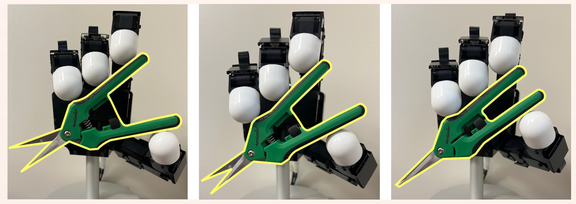

Fig. 9: The sequence from left to right illustrates the Allegro hand squeezing an articulated object. This example is manually generated and does not represent the outcome of a trained model.

References

[1] Chen, Sirui, Jeannette Bohg, and Karen Liu. "SpringGrasp: Synthesizing Compliant, Dexterous Grasps under Shape Uncertainty." 2nd Workshop on Dexterous Manipulation: Design, Perception and Control (RSS). 2024.
[2] Hadjivelichkov, Denis, et al. "One-shot transfer of affordance regions? AffCorrs!." Conference on Robot Learning. PMLR, 2023.
[3] Wang, Jun, et al. "Lessons from Learning to Spin" Pens"." arXiv preprint arXiv:2407.18902 (2024).
[4] Agarwal, Ananye, et al. "Dexterous functional grasping." 7th Annual Conference on Robot Learning. 2023.
[5] Katz, Dov. Interactive perception of articulated objects for autonomous manipulation. University of Massachusetts Amherst, 2011.

Acknowledgements

This work was conducted under the supervision of Jeannette Bohg and was funded by the summer CS research program at Stanford University. I would like to thank my mentor, Claire Chen, for her advice and guidance.

For more information, you can reach me at jk84@rice.edu or visit my LinkedIn page via the QR code.