3D Point Cloud Affordance Detection

Method Level Presentation

Lasse Huber-Saffer

Master Seminar Intelligent Robotics
Technical Aspects of Multimodal Systems
Department of Informatics
University of Hamburg

June 5, 2025

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Table of Contents

3D Affordance Detection

- Complex manipulation tasks require fine-grained object understanding
- 3D Affordance Detection Task
  - Input: 3D representation of objects/environment
    - (+ Textual Prompt)
  - Output: Affordance labels for individual regions
- Affordance vocabulary types
  - Closed ⇒ Predefined set of labels
  - Open ⇒ Ranging from unseen labels to complex freeform instructions

Figure: Example affordance labels[1]

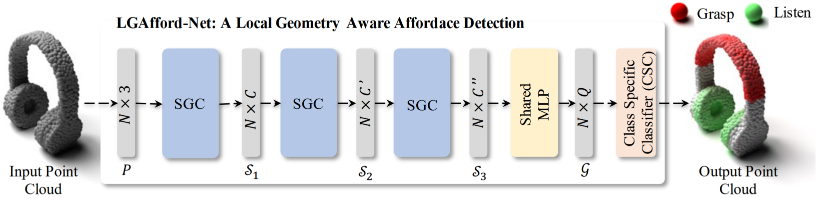
[1] Nguyen et al., IROS 2023

Overview

- Legacy approaches offer limited generalizability
  - Affordance vocabulary
  - Object types
- Recent approaches leverage pre-trained foundation models
  - Task-oriented scene understanding
  - Open-set textual task understanding
  - Object part segmentation
  - Grasp pose candidate generation
- Challenges
  - Affordances depend on robot setup
  - Incomplete information

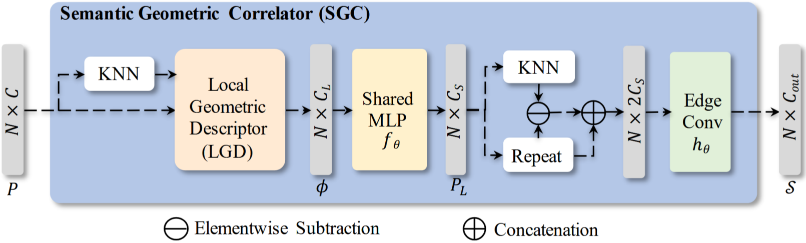
LGAfford-Net[2]

- Closed-vocabulary approach
- Trained on 3D AffordanceNet dataset
- Emphasis on local geometry



[2] Tabib et al., CVPR 2024

Semantic Geometric Correlator



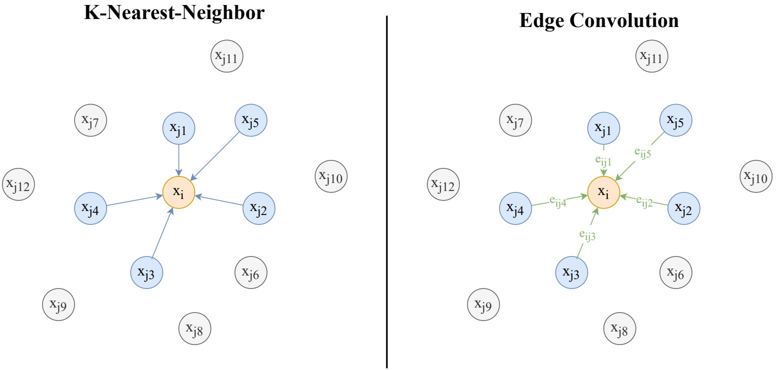
- P - Input Point Cloud
- ϕ - Local Geometric Features
- P_L - Learned Local Geometric Features
- $\mathcal{S}$ - Semantic Local Geometric Features

Local Geometric Descriptor

- Create triangles from each point $p_i \in \mathbb{R}^n$ and two nearest neighbors p_{j_1}, p_{j_2}
- Gather into higher-dimensional descriptor vector:

$$\phi_i = \begin{cases} p_i \\ \overrightarrow{e_1} = p_{j_1} - p_i \\ \overrightarrow{e_2} = p_{j_2} - p_i \\ |\overrightarrow{e_1}| \\ |\overrightarrow{e_2}| \\ \hat{n} = \overrightarrow{e_1} \times \overrightarrow{e_2} \\ \mu_i = mean(p_j) \\ \sigma_i = std(p_j) \end{cases}$$

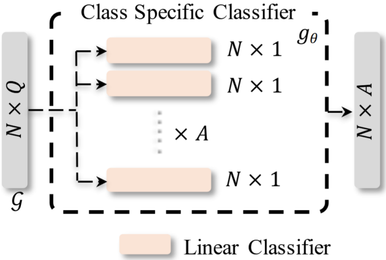
Edge Convolution[3]



- $\mathcal{S}$ - Semantic Local Geometric Features
  - $\mathcal{S}_i = \max_{j \in K_S}[h_\theta(x_j - x_i, x_i)]$

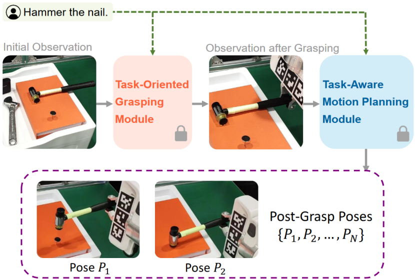
[3]Wang et al., ACM TOG 2019

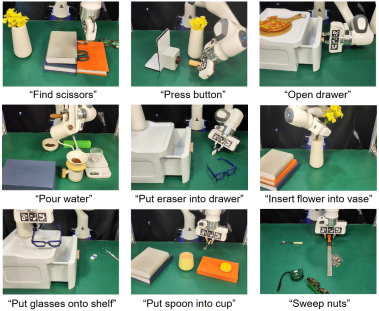
Class-Specific Classifier



- Train independent classifiers per affordance category
- Output for each point: Probability scores of all affordance categories

CoPa[4]

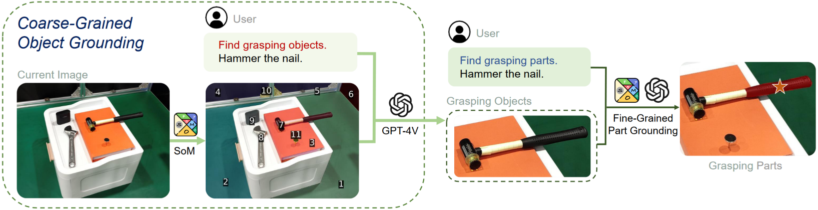




- Open-vocabulary approach
  - Textual task specification
- Utilize foundation models for affordance detection
- Outputs sequence of 6-DoF poses

[4] Huang et al., IROS 2024

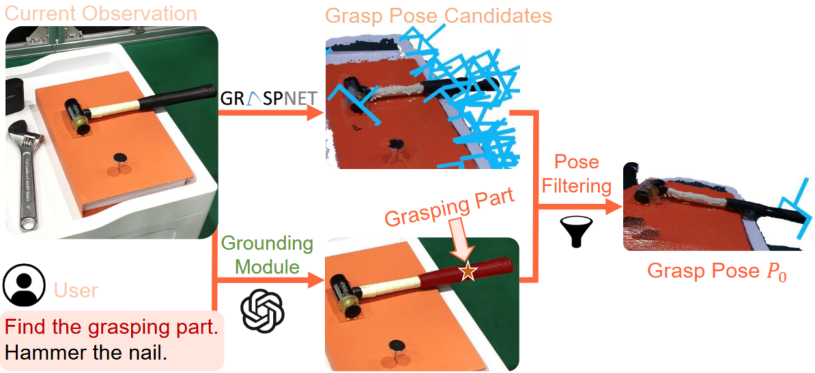
Grounding Module



- Set-of-Mark (SoM)[5] prompting for VLM
  - Segment image and assign numeric markers
- Coarse: Find graspable object in scene
- Fine: Find graspable part of object

[5] Yang et al., arXiv, 2023

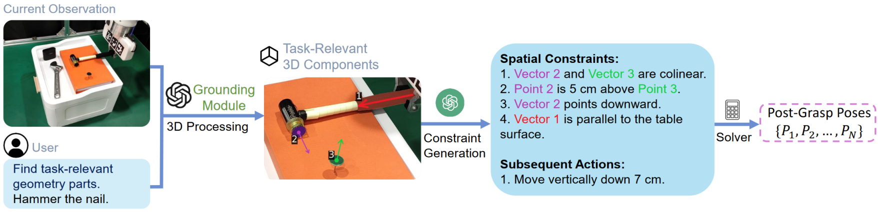
Task-Oriented Grasping Module



- Pose candidate generation using GraspNet[6]
- Select highest-confidence candidate within grasping part mask

[6] Fang et al., CVPR 2020

Task-Aware Motion Planning Module



- Task-relevant part grounding
- Manipulation constraint generation
  - Overlay simplified geometric indicators (vectors, surfaces)
  - Describe spatial constraints and action sequence using VLM
- Target pose planning
  - Nonlinear constraint solver
  - Sequentially compute post-grasp poses

Demo

Summary

- Emergence of foundation models has strongly impacted 3D affordance detection
- CoPa can easily be integrated into higher-level planning frameworks
  - Perform complex multi-step tasks
  - Explore unknown information
  - Adapt to dynamic scene changes

Limitations & Future Work

- Current challenges and limitations
  - Vector/surface constraint modeling is not sufficient for all object types and manipulation tasks
  - Spatial reasoning capabilities of VLMs are limited because of a lack of representation in input data
- Future research directions
  - Refine geometric constraint modeling
  - Incorporate RGB-D data into VLM training