Synergies Between Affordance and Geometry: 6-DoF Grasp Detection via Implicit Representations

Presenter: Kevin Black

9/9/2021
Robot Grasping
Robot Grasping: Considerations

- Geometric vs. data-driven
- Object model: known vs. unknown
- Sensor data:
  - RGB vs. RGB-D
  - Single-view vs. multi-view
- Open-loop vs. closed-loop
- Human-supervised vs. self-supervised learning
Robot Grasping: Considerations

- Geometric vs. **data-driven**
- Object model: known vs. **unknown**
- Sensor data:
  - RGB vs. RGB-D
  - **Single-view** vs. multi-view
- **Open-loop** vs. closed-loop
- Human-supervised vs. **self-supervised**
Prior Work

- Efficient grasping from RGBD images: Learning using a new rectangle representation (ICRA 2011)
  - Introduces the hand-labeled Cornell grasping dataset, which became widely used
- Dex-Net 2.0: Deep Learning to Plan Robust Grasps with Synthetic Point Clouds and Analytic Grasp Metrics (RSS 2017)
  - A continuing line of work that has explored multiple gipper types (e.g. suction)
- Deep Object Pose Estimation for Semantic Robotic Grasping of Household Objects (CoRL 2018)
  - Performs pose estimation on known objects in order to geometrically compute grasps
- Grasping in the Wild: Learning 6DoF Closed-Loop Grasping from Low-Cost Demonstrations (ICRA 2020)
  - Learns from real human demonstrations, uses reinforcement learning for closed-loop feedback
Problem Formulation

Input: single-view RGB-D image

Output: grasp affordance

\[
\begin{aligned}
    t & \in \mathbb{R}^3 & \text{Grasp center} \\
    w & \in [0, w_{max}] & \text{Gripper width} \\
    r & \in SO(3) & \text{Gripper rotation} \\
    q & \in [0, 1] & \text{Grasp quality} \\
    \end{aligned}
\]

Performance measured using grasp success rate
Target Scenarios

Packed objects (more occlusion)  Piled objects (less occlusion)
Primary Contribution

Grasp Detection via Implicit Geometry and Affordance (GIGA)

- Takes a single side view RGB-D image as input
- Network architecture which utilizes implicit neural representations
- Also learns 3D reconstruction as an auxiliary task

- Key insight: learning grasp detection and 3D reconstruction in tandem will synergize to improve overall performance in both tasks
- 3D reconstruction will especially help to reason about occluded regions
Truncated Signed Distance Function (TSDF)

[Werner et al. 2014]
Implicit Neural Representations

Represent data as a function from spatial coordinates to values, i.e.

$$f : \mathbb{R}^n \rightarrow \mathcal{Y}$$

Usually conditioned on additional input:

$$f : \mathbb{R}^n \times \mathcal{X} \rightarrow \mathcal{Y}$$

Advantages:
- “Infinite resolution” (no discretization)
- Memory-efficient

[Skorokhodov et al. 2020]
Implicit Representations for 3D Reconstruction

Voxels  Point cloud  Mesh  Implicit occupancy function

[Mescheder et al. 2019]
GIGA: Grasp Detection + 3D Reconstruction

Grasp detection:

\[ f_a : t, x \rightarrow q, r, w \]

3D reconstruction:

\[ f_g : p, x \rightarrow b \]

Candidate grasp center (3D point)\nGrasp quality, rotation, width\nInput scene

Any 3D point \nOccupancy value
GIGA Architecture

*Backbone architecture largely based on prior work: Convolutional Occupancy Networks (2020) by Peng et al.
GIGA Architecture

Projected 2D feature grids

Structured feature grids

2D U-Nets

*Backbone architecture largely based on prior work: Convolutional Occupancy Networks (2020) by Peng et al.
GIGA Architecture

*Backbone architecture largely based on prior work: Convolutional Occupancy Networks (2020) by Peng et al.*
Training Details

- Trained in simulation
- Grasps randomly sampled around surface of objects
- Grasp width and rotation only trained on successful grasps
- Data must be balanced by eliminating extra unsuccessful grasps
- 3D reconstruction trained using uniformly sampled points
- Noise added to images to aid with sim2real transfer
Closest Existing Approach: Volumetric Grasping Network (VGN)

- Predicts grasp affordance for each voxel rather than using an implicit neural representation
- Does not learn 3D reconstruction

[Breyer et al. 2021]
## GIGA Grasping Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Packed</th>
<th></th>
<th>Pile</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GSR (%)</td>
<td>DR (%)</td>
<td>GSR (%)</td>
<td>DR (%)</td>
</tr>
<tr>
<td>SHAF [13]</td>
<td>56.6 ± 2.0</td>
<td>58.0 ± 3.0</td>
<td>50.7 ± 1.7</td>
<td>42.6 ± 2.8</td>
</tr>
<tr>
<td>GPD [16]</td>
<td>35.4 ± 1.9</td>
<td>30.7 ± 2.0</td>
<td>17.7 ± 2.3</td>
<td>9.2 ± 1.3</td>
</tr>
<tr>
<td>VGN [4]</td>
<td>74.5 ± 1.3</td>
<td>79.2 ± 2.3</td>
<td>60.7 ± 4.2</td>
<td>44.0 ± 4.9</td>
</tr>
<tr>
<td>GIGA-Aff</td>
<td>77.2 ± 2.3</td>
<td>78.9 ± 1.7</td>
<td>67.8 ± 3.0</td>
<td>49.7 ± 1.9</td>
</tr>
<tr>
<td>GIGA</td>
<td>83.5 ± 2.4</td>
<td>84.3 ± 2.2</td>
<td>69.3 ± 3.3</td>
<td>49.8 ± 3.9</td>
</tr>
<tr>
<td>GIGA (HR)</td>
<td>87.9 ± 3.0</td>
<td>86.0 ± 3.2</td>
<td>69.8 ± 3.2</td>
<td>51.1 ± 2.8</td>
</tr>
</tbody>
</table>

**GIGA-Aff:** without the reconstruction component  
**GIGA (HR):** high resolution (60x60x60 sampled grasp candidates rather than 40x40x40)  
**GSR:** grasp success rate  
**DR:** declutter rate (proportion of items removed after running until failure)
GIGA Grasping Results
### GIGA 3D Reconstruction Results

<table>
<thead>
<tr>
<th>Method</th>
<th>IoU (%)</th>
<th>IoU-Grasp (%)</th>
<th>Δ% (IoU-Grasp—IoU)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GIGA-Detach</td>
<td>53.2</td>
<td>68.8</td>
<td>+15.6</td>
</tr>
<tr>
<td>GIGA</td>
<td>70.0</td>
<td>78.1</td>
<td>+8.1</td>
</tr>
<tr>
<td>GIGA-Geo</td>
<td>80.0</td>
<td>84.0</td>
<td>+4.0</td>
</tr>
</tbody>
</table>

**GIGA-Detach**: features trained for grasping, weights frozen, final layers trained for reconstruction  

**IoU**: intersection-over-union of reconstructed object  

**IoU-Grasp**: the IoU around graspable regions only  

**GIGA-Geo**: end-to-end trained for reconstruction only
GIGA 3D Reconstruction Results
# Real Robot Experiments

<table>
<thead>
<tr>
<th>Method</th>
<th>Packed GSR (%)</th>
<th>DR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGN [4]</td>
<td>77.2 (61/79)</td>
<td>81.3</td>
</tr>
<tr>
<td>GIGA</td>
<td><strong>83.3</strong> (65/78)</td>
<td><strong>86.6</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Pile GSR (%)</th>
<th>DR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGN [4]</td>
<td>79.0 (64/81)</td>
<td>85.3</td>
</tr>
<tr>
<td>GIGA</td>
<td><strong>86.9</strong> (73/84)</td>
<td><strong>97.3</strong></td>
</tr>
</tbody>
</table>
Limitations/Future Work

- Currently throws away reconstructed 3D information at test time
  - Could be used for closed-loop control

- Assumes a single static viewpoint
  - Not generalizable to a mobile robot or camera system

- sim2real transfer not very thoroughly evaluated
  - A real out-of-laboratory setting would be much more noisy and likely hurt performance significantly

- Could possibly be extended to other manipulation tasks?
Summary

- Grasping arbitrary objects is a very hard but fundamentally useful task
- GIGA does not rely on known object models, multiple views, or uncluttered/unoccluded scenes
- Key insights:
  - Implicit neural representations work well for efficiently representing grasp affordance
  - Learning 3D reconstruction synergizes with grasping, especially for occluded objects
- GIGA demonstrates state-of-the-art results on cluttered grasping from a single view
Thank you!
Extended Reading

- A Survey on Learning-Based Robotic Grasping
- Volumetric Grasping Network
- Occupancy Networks
- Convolutional Occupancy Networks