Connecting 6D Object Pose Estimation with Robot Manipulation

Yu Xiang
Assistant Professor
Computer Science
The University of Texas at Dallas

10/3/2023
8th International Workshop on Recovering 6D Object Pose @ ICCV 2023
Model-based 6D Object Pose Estimation

- Camera
- 3D world
- 3D rotation \( R \)
- 3D Translation \( T = (T_x, T_y, T_z)^T \)
- Object coordinate

3D models
Model-based 6D Object Pose Estimation

• What information can be obtained from 6D object pose estimation?
  • Object position in camera frame \( T = (T_x, T_y, T_z)^T \)
  • Object orientation in camera frame \( R \)

Input image  
Point cloud and object axes in camera frame  
Projection of 3D models onto the input image
Applications

Robot Manipulation

Augmented Reality

Industrial Object Inspection

https://www.mvtec.com/company/research/datasets/mvtec-itodd
Model-Based Robot Manipulation

Perception
6D object pose estimation

Planning
Grasp planning and motion planning

Control
Manipulation trajectory following

2X
Traditional Methods for 6D Object Pose Estimation

- Feature matching-based methods

Rothganger-Lazebnik-Schmid-Ponce, IJCV’06

Collet-Martinez-Srinivasa, IJRR’11

- Template matching-based methods

Hinterstoisser-Lepetit-Ilic-Holzer-Bradski-Konolige-Navab, ACCV’12

Choi-Christensen, IROS’12
PoseCNN: the First End-to-end 6D Pose Estimation Network

✓ Texture-less objects
✓ Symmetric objects
✓ Occlusions

Xiang-Schmidt-Narayanan-Fox, RSS’18
PoseCNN: the First End-to-end 6D Pose Estimation Network

Input image

Feature extraction

3D translation estimation

Semantic segmentation

Center direction X

Center direction Y

Center distance

3D Translation estimation

\[ \mathbf{T} = (T_x, T_y, T_z)^T \]

For each RoI

3D Rotation regression

For each RoI

Hough voting layer

6D Poses

Xiang-Schmidt-Narayanan-Fox, RSS’18
PoseCNN: the First End-to-end 6D Pose Estimation Network

Segmentation and Detection

Poses

3D World

Input image

Xiang-Schmidt-Narayanan-Fox, RSS’18
DeepIM: Deep Iterative Matching for 6D Pose Estimation

Li-Wang-Ji-Xiang-Fox, ECCV’18 Oral, IJCV’19

CosyPose: Labbe-Carpentier-Aubry-Sivic, ECCV’20
DeepIM: Deep Iterative Matching for 6D Pose Estimation

<table>
<thead>
<tr>
<th></th>
<th>RGB</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>YCB Video</td>
<td>PoseCNN</td>
<td>PoseCNN+ DeepIM</td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>75.9</td>
<td><strong>88.1</strong></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>RGB-D</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>YCB Video</td>
<td>PoseCNN+ ICP</td>
<td>PoseCNN+ DeepIM</td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>93.0</td>
<td><strong>94.0</strong></td>
<td></td>
</tr>
</tbody>
</table>

Li-Wang-Ji-Xiang-Fox, ECCV’18 Oral, IJCV’19
PoseRBPF: A Rao-Blackwellized Particle Filter for 6D Object Pose Tracking

A particle $\mathcal{X}_i = \{\mathbf{T}_i, P(\mathbf{R}_i | \mathbf{T}_i, \mathbf{Z}_{1:k})\}$

3D Translation $\mathbf{T}_i$

Orientation Distribution $P(\mathbf{R}_i | \mathbf{T}_i, \mathbf{Z}_{1:k})$

191,808 bins

Codebook

Encoder

Particle Code

Rotation Likelihood

Input frame

Rol

Sundermeyer-Marton-Durner-Brucker-Triebel, ECCV’18

Deng-Mousavian-Xiang-Xia-Bretl-Fox, RSS’19, T-RO’21
PoseRBPF: A Rao-Blackwellized Particle Filter for 6D Object Pose Tracking

Overlay of 3D Models  2D Segmentation  Input Image

3D World

With SDF-based depth refinement
GDR-Net: Geometry-guided Direct Regression Network

Overall best method in the BOP Challenge 2022

Model-Based Robot Manipulation

Perception
- 6D object pose estimation

Planning
- Grasp planning and motion planning

Control
- Manipulation trajectory following
Grasp Planning: GraspIt!

GraspIt!  [https://graspit-simulator.github.io/](https://graspit-simulator.github.io/)
Grasp Planning: A Physics-based Approach
Motion Planning

The Open Motion Planning Library in MoveIt

https://ompl.kavrakilab.org/index.html

18
OMG Planner: Trajectory Optimization and Grasp Selection

Modeling the goal set distribution

OMG Iter: 50

100 grasps

Code available online
Model-Based Robot Manipulation

Perception
6D object pose estimation

Planning
Grasp planning and motion planning

Control
Manipulation trajectory following

20
Benchmarks

- 6D object pose
Benchmarking

- Robot manipulation

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>Type</th>
<th>Task</th>
<th>Objects</th>
<th>AR Tag-Free</th>
<th>Scene Reproducibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>RLBench [12]</td>
<td>Simulation</td>
<td>100 Tasks</td>
<td>Synthetic</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>robosuite [13]</td>
<td>Simulation</td>
<td>9 Tasks</td>
<td>Synthetic</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>NIST Assembly [7]</td>
<td>Real</td>
<td>Assembly</td>
<td>Task Boards</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>FurnitureBench [14]</td>
<td>Real</td>
<td>Assembly</td>
<td>3D Printing</td>
<td>X</td>
<td>✓</td>
</tr>
<tr>
<td>GRASPA [8]</td>
<td>Real</td>
<td>Grasping</td>
<td>YCB (clutter)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>OCRTOC [15]</td>
<td>Real</td>
<td>Rearrangement</td>
<td>YCB + Others</td>
<td>✓</td>
<td>X</td>
</tr>
<tr>
<td>RB2 [16]</td>
<td>Real</td>
<td>Pouring, Scooping, Zipping, Insertion</td>
<td>Others</td>
<td>✓</td>
<td>X</td>
</tr>
<tr>
<td>Box and Blocks Test [17]</td>
<td>Real</td>
<td>Pick-and-Place</td>
<td>Blocks</td>
<td>✓</td>
<td>X</td>
</tr>
<tr>
<td><strong>SceneReplica (Ours)</strong></td>
<td>Real</td>
<td>Pick-and-Place</td>
<td>YCB (clutter)</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

NIST Assembly board  FurnitureBench  GRASPA  SceneReplica
SceneReplica for Real-World Robot Manipulation

- 16 YCB objects
- Stable poses
- Reachability testing in simulation
SceneReplica for Real-World Robot Manipulation

20 Scenes

SceneReplica: https://irvlutd.github.io/SceneReplica/
Real-World Scene Setup

Reference Image

Real World Setup
Model-based Grasping vs Model-free Grasping

6D Pose Estimation  Offline Grasp Database  Motion Planning Setup  Grasping & Lifting  Moving arm for Dropoff

Input real world scene

Unseen Object Segmentation  Model-free Grasp Planning  Motion Planning Setup  Grasping & Lifting  Moving arm for Dropoff
Model-based Grasping Example
Model-free Grasping Example

SceneReplica Benchmark
MSMFormer | Contact GraspNet + Top Down | Movelt
Scene: 130 | Order: Random

Rviz Capture
Realsense Capture
End-to-end Learning-based Grasping

Dex-Net 2.0 Grasping Example
## Current Leaderboard

<table>
<thead>
<tr>
<th>#</th>
<th>Perception</th>
<th>Grasp Planning</th>
<th>Motion Planning</th>
<th>Control</th>
<th>Ordering</th>
<th>Grasping Type</th>
<th>Pick &amp; Place Success</th>
<th>Grasping Success</th>
<th>Videos</th>
</tr>
</thead>
</table>

Overall best method in the BOP Challenge 2022

https://irvlutd.github.io/SceneReplica/
Failure Analysis (GDRNPP + GraspIt! + Top-Down)

Object Detection Error

Pose Estimation Error

Grasp Planning Error

<table>
<thead>
<tr>
<th>Object</th>
<th>S</th>
<th>$P_F$</th>
<th>$P_{EF}$</th>
<th>EF</th>
</tr>
</thead>
<tbody>
<tr>
<td>003 cracker box</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>004 sugar box</td>
<td>5</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>005 tomato soup can</td>
<td>5</td>
<td>1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>006 mustard bottle</td>
<td>7</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>007 tuna fish can</td>
<td>1</td>
<td>5</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>008 pudding box</td>
<td>5</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>009 gelatin box</td>
<td>6</td>
<td>-</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>010 potted meat can</td>
<td>7</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>011 banana</td>
<td>6</td>
<td>-</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>021 bleach cleanser</td>
<td>3</td>
<td>1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>024 bowl</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>025 mug</td>
<td>4</td>
<td>-</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>037 scissors</td>
<td>4</td>
<td>3</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>035 power drill</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>040 large marker</td>
<td>2</td>
<td>4</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>055 extra large clamp</td>
<td>5</td>
<td>-</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td><strong>ALL</strong></td>
<td>66</td>
<td>24</td>
<td>7</td>
<td>3</td>
</tr>
</tbody>
</table>

$P_{EF}$: #perception failure

$P_{LF}$: #planning failure

EF: #execution failure
Conclusion

- 6D object pose estimation can facilitate robot manipulation
- The performance of 6D object pose estimation is not saturated yet
- Connecting BOP and SceneReplica to evaluate object pose estimation and robot grasping in the real world

Thank you!