Object Pose Estimation

Tom Sanitz

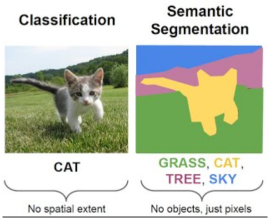
16. November 2013

Table of Contents

Related topics

2D image domain

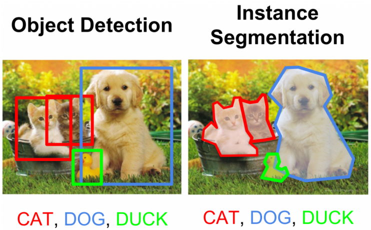
1. Classification
   - Many advances in last 10 years (CNNs / Transformers)
   - Important for many downstream tasks
2. Semantic Segmentation
   - Similar to image classification
   - Per pixel classification



Related topics

2D Multi Object

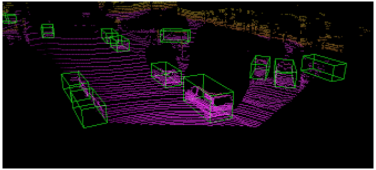
1. Object Detection
   - 2D localisation of multiple objects in (rotated) bounding boxes + class id
   - Examples: YOLO [5], Faster R-CNN [6], ...
2. Instance Segmentation
   - Hybrid between Object Detection and Semantic Segmentation
   - Example: Mask R-CNN [3]



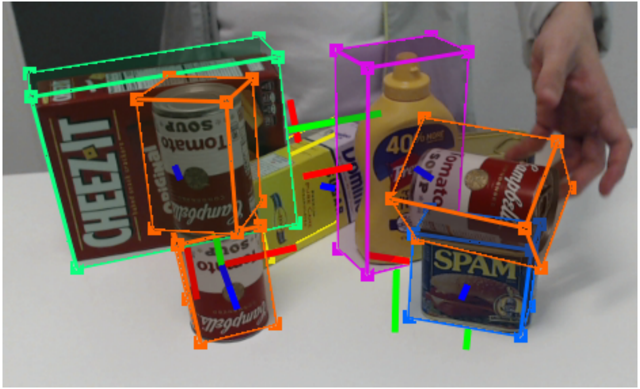
Related topics

3 dimensional domain

1. 3D Object Detection
   - Predicts 3d center of the box, width, height and length (+ vertical rotation)



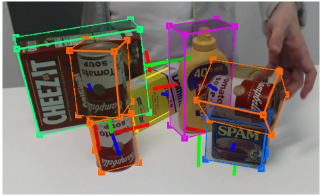
6D Pose Estimation



6D Pose Estimation

What do we want to achieve

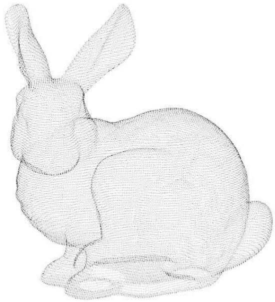
- Prediction of 3D position + 3D rotation
- But 3D position of what? Relative to where?



6D Pose Estimation

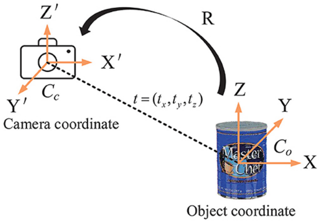
Defining a coordinate system

- 3D Representation of the object (e.g pointcloud)
- Used to define object intrinsic pose



6D Pose Estimation

- Representation of object pose in camera frame
- Finding transformation of objects frame to camera frame, see robotics lecture



Synthetic Dataset generation

Synthetic Datasets

Allows the creation of massive amounts of training data at minimal cost.

- When more data is required
- When real data is to expensive to annotate

Problems with synthetic data

Potentially a bad representation of real data: Reality gap

Deep Object Pose Estimation (DOPE)

- Deep Object Pose Estimation for Semantic Robotic Grasping of Household Objects [7]
- Published at "Conference on Robot Learning (CoRL) 2018" by Tremblay et al.(NVIDIA)
- Simple network architecture
- Trained only on synthetic data
- Achieved state-of-the-art performance

Synthetic Dataset generation

Reality gap solutions in DOPE

- Photo realistic renders in UE 4
- Domain randomization

Synthetic Dataset generation

Domain Randomization

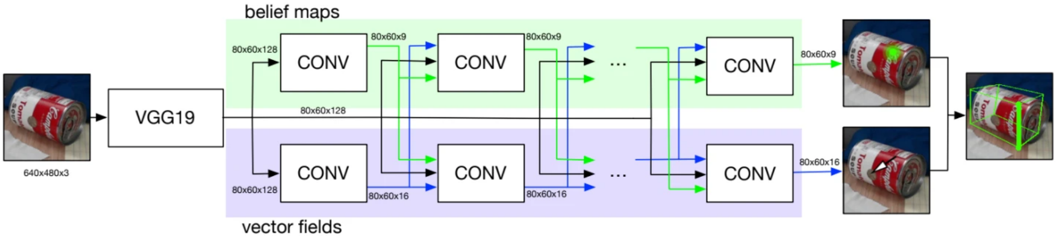
- Object of interest in front of random backgrounds
- Adding various distractors
- randomize and overlay textures, lighting, poses and noise
- Force the network to learn general features

domain randomized

Network architecture

DOPE [7] - Fully convolutional Network architecture

1. Predict 9 belief maps of 2D keypoints per objects
2. Learn 8 vector fields for vertices-centroid assignment
3. Use perspective-n-point (PnP) on the peaks to estimate 6D Pose



Common Objects

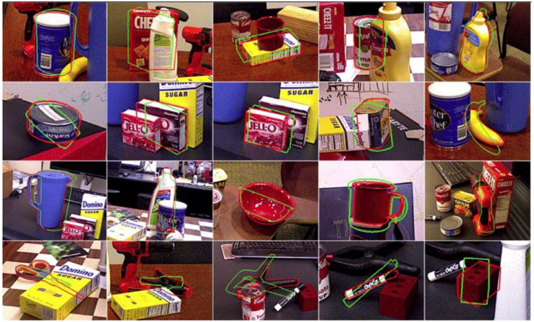
YCB Objects

- Common household objects
- Meant to standardise evaluation methods between research
- Different materials and shapes
- Also used in the YCB-Video dataset [8]

Evaluation Dataset

YCB Video Dataset [8]

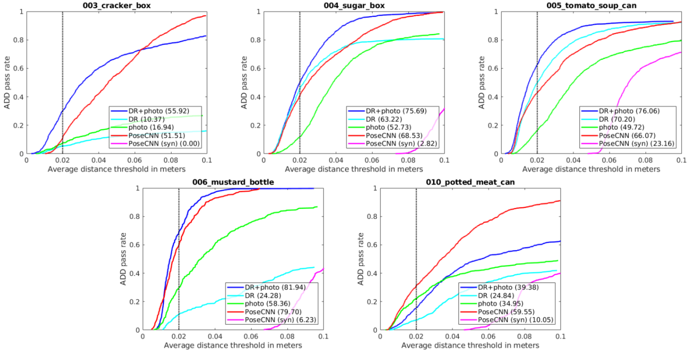
- Dataset including accurate 6D poses for YCB objects
- Originally 21 YCB objects included
- 92 Videos with 133.827 frames



Evaluation Metrics

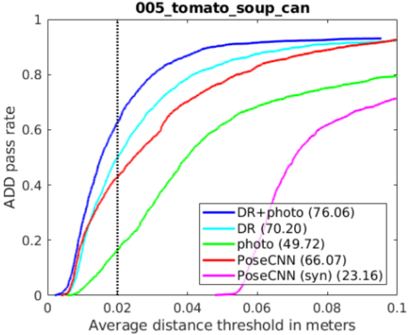
Average Distance (ADD)

- Average 3D Euclidean distance of all model points
- ADD pass rate = Percentage of predictions(p) with ADD value <= threshold (t)



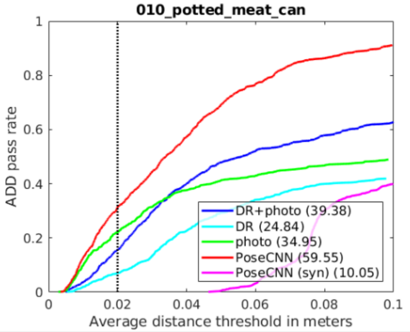
Evaluation Metrics

- Note the difference between synthetic and domain randomization



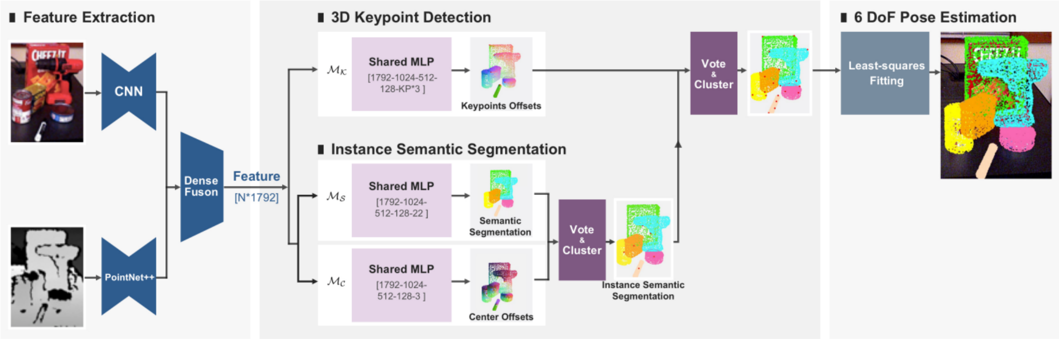
Evaluation Metrics

- Why does DOPE perform worse than PoseCNN [8] here?
- Answer: Reality gap for metallic objects



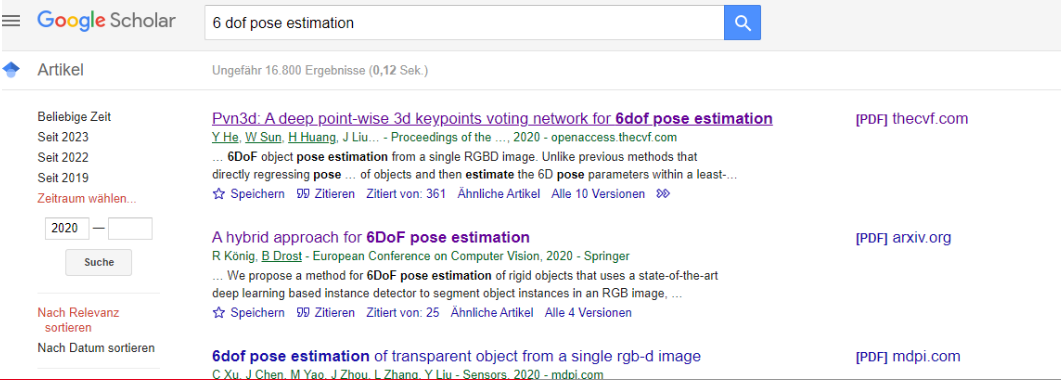
RGB-D Based Solutions

- What if we include depth information?
- PVN3D: A Deep Point-Wise 3D Keypoints Voting Network [4]



And many more..

- EfficientPose [1]
- RNN Pose [9]
- ROPE [2]
- ...



DOPE Demo

- Showcasing live DOPE demo
- Inference on the YCB tomato soup object
- Using rgb only on the intel realsense D435 camera

Thank you for your attention!

References

[1] Yannick Bukschat and Marcus Vetter. "EfficientPose: An efficient, accurate and scalable end-to-end 6D multi object pose estimation approach". In: arXiv preprint arXiv:2011.04307 (2020).

[2] Bo Chen, Tat-Jun Chin, and Marius Klimavicius. "Occlusion-robust object pose estimation with holistic representation". In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2022.

[3] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. "Mask r-cnn". In: Proceedings of the IEEE international conference on computer vision. 2017.

[4] Yisheng He, Wei Sun, Haibin Huang, Jianran Liu, Haoqiang Fan, and Jian Sun. "Pvn3d: A deep point-wise 3d keypoints voting network for 6dof pose estimation". In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.

[5] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. "You only look once: Unified, real-time object detection". In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[6] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. "Faster r-cnn: Towards real-time object detection with region proposal networks". In: Advances in neural information processing systems (2015).

[7] Jonathan Tremblay, Thang To, Balakumar Sundaralingam, Yu Xiang, Dieter Fox, and Stan Birchfield. "Deep Object Pose Estimation for Semantic Robotic Grasping of Household Objects". In: Conference on Robot Learning (CoRL). 2018. URL: https://arxiv.org/abs/1809.10790.

[8] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes". In: arXiv preprint arXiv:1711.00199 (2017).

[9] Yan Xu, Kwan-Yee Lin, Guofeng Zhang, Xiaogang Wang, and Hongsheng Li. "Rnnpose: Recurrent 6-dof object pose refinement with robust correspondence field estimation and pose optimization". In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

ADD vs. ADD-S

ADD

- R: Ground truth rotation
- T: Ground truth translation
- $\tilde{T}$ and $\tilde{R}$ the predicted values
- M are the set of 3D model points, m the number of points
- Compute the mean of pairwise distances

$$\text{ADD} = \frac{1}{m} \sum_{\mathbf{x} \in \mathcal{M}} \|(\mathbf{R}\mathbf{x} + \mathbf{T}) - (\tilde{\mathbf{R}}\mathbf{x} + \tilde{\mathbf{T}})\|$$

ADD vs. ADD-S

ADD-S

- For symmetric objects, matching can be ambiguous
- Solution in DOPE: closest point distance

$$\mathrm{ADD} = \frac{1}{m} \sum_{\mathbf{x} \in \mathcal{M}} \|(\mathbf{R}\mathbf{x} + \mathbf{T}) - (\tilde{\mathbf{R}}\mathbf{x} + \tilde{\mathbf{T}})\|$$