Multiview RGB-D Dataset for Object Instance Detection
Georgios Georgakis, Md. Alimoor Reza, Arsalan Mousavian, Phi-Hung Le and Jana Kosecka
George Mason University

Overview

Contributions:
• A new RGB-D dataset of cluttered kitchen scenes, annotated in both 2D and 3D, for detection and recognition of hand-held objects in realistic settings. Some objects were taken from the BigBird dataset [1].
URL: http://cs.gmu.edu/~robot/gmu-kitchens.html

• A multiview object proposal generation method which uses only 3D information.

• Detection baselines that investigate how different training strategies can affect the performance of CNNs.

Multiview Object Proposals

Steps:
1) Removal of large planar surfaces from the dense point cloud.
2) Mean-shift clustering of remaining points in multiple ranges.
3) Cuboid fitting for removing outlier points.

Object Detection

Baselines training:
1) Turntable: Cropped object images from BigBird[1].
2) Turntable background: Same as (1) augmented with images superimposed on random backgrounds.
3) HMP Folds: Scenes are split into three training-test sets and HMP[2] is used.
4) CNN Folds: Same as (3) but we train a CNN instead of HMP. Baselines (1),(2),(4), train a CNN.

Kitchen Scenes Dataset

Procedure:
• Collected the scenes with Kinect V2 (1920x1080).
• Sparse reconstructions are created with the latest structure from motion (SFM) software COLMAP.
• Dense point clouds are created using the estimated camera poses to project all points to the world coordinate frame.

Contents:
• 9 RGB-D kitchen video sequences (6735 images).
• 10-15 object instances per scene, with 23 instances in total.
• Bounding box annotations for all objects.
• 3D point labeling for each scene.

Comparison to WRGB-D[2]

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Recall(%)</th>
<th>No. Proposals</th>
</tr>
</thead>
<tbody>
<tr>
<td>WRGB-D[2]</td>
<td>86.3/87</td>
<td></td>
</tr>
<tr>
<td>Our Kitchen Scenes</td>
<td>62.6/2999</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Performance of the single-view proposal generation algorithms on the WRGB-D[2] and our Kitchen Scenes dataset.

Conclusions

• Multiview 3D object proposals outperform singleview 3D proposals and are comparable to established proposal techniques.
• Training on similar backgrounds as the test set leads to much better performing detectors, however that data are hard to acquire. Training on random backgrounds helps just slightly, which suggests that more sophisticated approaches are needed.
• Comparative experiments on the WRGB-D [2] show that the Kitchen scenes dataset is more challenging.

References

Acknowledgments: We acknowledge support from NSF NRI grant 1527208.