Semantic RGB-D Perception for Cognitive Robots

Sven Behnke

Computer Science Institute VI
Autonomous Intelligent Systems
Our Domestic Service Robots

Dynamaid
- Size: 100-180 cm, weight: 30-35 kg
- 36 articulated joints
- PC, laser scanners, **Kinect**, microphone, ...

Cosero
RoboCup 2013 Eindhoven
Analysis of Table-top Scenes and Grasp Planning

- Detection of clusters above horizontal plane
- Two grasps (top, side)

- Flexible grasping of many unknown objects

[Stückler, Steffens, Holz, Behnke, Robotics and Autonomous Systems 2012]
3D-Mapping with Surfels
3D-Mapping with Surfels
3D-Mapping and Localization

- Registration of 3D laser scans
- Representation of point distributions in voxels
- Drivability assessment through region growing
- Robust localization using 2D laser scans
3D Mapping by RGB-D SLAM

- Modelling of shape and color distributions in voxels
- Local multiresolution
- Efficient registration of views on CPU
- Global optimization
- Multi-camera SLAM

[Stückler, Behnke: Journal of Visual Communication and Image Representation 2013]

[Stoucken, Diplomarbeit 2013]
Learning and Tracking Object Models

- Modeling of objects by RGB-D-SLAM

- Real-time registration with current RGB-D image
Deformable RGB-D-Registration

- Based on Coherent Point Drift method [Myronenko & Song, PAMI 2010]
- Multiresolution Surfel Map allows real-time registration

[Stückler, Behnke, ICRA2014]
Transformation of Poses on Object

- Derived from the deformation field

[Stückler, Behnke, ICRA2014]
Grasp & Motion Skill Transfer

- Demonstration at RoboCup 2013 [Stückler, Behnke, ICRA2014]
Tool use: Bottle Opener

- Tool tip perception
- Extension of arm kinematics
- Perception of crown cap
- Motion adaptation

[Stückler, Behnke, Humanoids 2014]
Picking Sausage, Bimanual Transport

- Perception of tool tip and sausage
- Alignment with main axis of sausage

- Our team NimbRo won the RoboCup@Home League in three consecutive years

[Stückler, Behnke, Humanoids 2014]
Hierarchical Object Discovery through Motion Segmentation

- Motion is a strong segmentation cue
- Both camera and object motion

Segment-wise registration of a sequence

Inference of a segment hierarchy

[Stückler, Behnke: IJCAI 2013]
Semantic Mapping

- Pixel-wise classification of RGB-D images by random forests
- Inner nodes compare color / depth of regions
- Size normalization
- Training and recall on GPU
- 3D fusion through RGB-D SLAM
- Evaluation on own data set and NYU depth v2

<table>
<thead>
<tr>
<th>Accuracy in %</th>
<th>Ø Classes</th>
<th>Ø Pixels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Silberman et al. 2012</td>
<td>59,6</td>
<td>58,6</td>
</tr>
<tr>
<td>Couprie et al. 2013</td>
<td>63,5</td>
<td>64,5</td>
</tr>
<tr>
<td>Random forest</td>
<td>65,0</td>
<td>68,1</td>
</tr>
<tr>
<td>3D-Fusion</td>
<td>66,8</td>
<td>70,6</td>
</tr>
</tbody>
</table>

[Stückler et al., Journal of Real-Time Image Processing 2014]
Learning Depth-sensitive CRFs

- SLIC+depth super pixels
- Unary features: random forest
- Height feature

Pairwise features
- Color contrast
- Vertical alignment
- Depth difference
- Normal differences

Results:
- Random forest
- CRF prediction
- Ground truth

<table>
<thead>
<tr>
<th></th>
<th>class average</th>
<th>pixel average</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF</td>
<td>65.0</td>
<td>68.3</td>
</tr>
<tr>
<td>RF + SP</td>
<td>65.7</td>
<td>70.1</td>
</tr>
<tr>
<td>RF + SP + SVM</td>
<td>70.4</td>
<td>70.3</td>
</tr>
<tr>
<td>RF + SP + CRF</td>
<td><strong>71.9</strong></td>
<td><strong>72.3</strong></td>
</tr>
<tr>
<td>Silberman et al.</td>
<td>59.6</td>
<td>58.6</td>
</tr>
<tr>
<td>Couprie et al.</td>
<td>63.5</td>
<td>64.5</td>
</tr>
</tbody>
</table>

[Müller and Behnke, ICRA 2014]
Object Class Detection in RGB-D

- Hough forests make not only object class decision, but describe object center
- RGB-D objects data set
- Color and depth features
- Training with rendered scenes
- Detection of object position and orientation

Depth helps a lot

[Badami, Stückler, Behnke: SPME 2013]
Bin Picking

- Known objects in transport box
- Matching of graphs of 2D and 3D shape primitives
- Grasp and motion planning

[Nieuwenhuisen et al.: ICRA 2013]
Learning of Object Models

- Scan multiple objects in different poses
- Find support plane and remove it
- Segment views
- Register views using ICP
- Recognize geometric primitives

Registered views  Surface reconstruction  Detected primitives
Active Object Perception

[Holz et al. STAR 2014]
Active Object Perception

- Efficient exploration of the part arrangement in the transport boxes to handle occlusions

[Holz et al. STAR 2014]
Active Object Perception

- Efficient exploration of the part arrangement in the transport boxes to handle occlusions
Efficient exploration of the part arrangement in the transport boxes to handle occlusions

[Holz et al. STAR 2014]
Industrial Application: Depalettizing

- Using work space RGB-D camera
- Initial pose of transport box roughly known
- Detect dominant horizontal plane above ground
- Cluster points above support plane
- Estimate main axes

[Holz et al. IROS 2015]
Object View Registration

- Wrist RGB-D camera moved above innermost object candidate
- Object views are represented as Multiresolution Surfel Map
- Registration of object view with current measurements using soft assignments
- Verification based on registration quality

[Holz et al. IROS 2015]
We detect potential object candidates using the workspace camera.

[Holz et al. IROS 2015]
Depalletizing Results: 10 Runs

- Total time

<table>
<thead>
<tr>
<th>Component</th>
<th>Mean</th>
<th>Std</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Object detection and grasping</td>
<td>13.84 s</td>
<td>1.89 s</td>
<td>10.42 s</td>
<td>23.81 s</td>
</tr>
<tr>
<td>Full cycle (incl. release and returning to initial pose)</td>
<td>34.57 s</td>
<td>3.01 s</td>
<td>29.53 s</td>
<td>49.52 s</td>
</tr>
</tbody>
</table>

- Component times and success rates

<table>
<thead>
<tr>
<th>Component</th>
<th>Mean</th>
<th>Std</th>
<th>Min</th>
<th>Max</th>
<th>Success Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial object detection</td>
<td>26.3 ms</td>
<td>10.3 ms</td>
<td>0.02 ms</td>
<td>38.5 ms</td>
<td>100 %</td>
</tr>
<tr>
<td>Detecting that the pallet is empty</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>100 %</td>
</tr>
<tr>
<td>Object localization &amp; verification</td>
<td>532.7 ms</td>
<td>98.2 ms</td>
<td>297.0 ms</td>
<td>800.1 ms</td>
<td>100 %</td>
</tr>
<tr>
<td>Identifying wrong objects</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>100 %</td>
</tr>
<tr>
<td>Grasping a found object</td>
<td>7.80 s</td>
<td>0.56 s</td>
<td>6.90 s</td>
<td>10.12 s</td>
<td>99 %</td>
</tr>
</tbody>
</table>

[Holz et al. IROS 2015]
Part Verification Results

- Parts used for verification

- Detection confidences

<table>
<thead>
<tr>
<th>Object</th>
<th>Mean</th>
<th>Std</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct object (“cross clamp”)</td>
<td>0.901</td>
<td>0.024</td>
<td>0.853</td>
<td>0.951</td>
</tr>
<tr>
<td>Similar cross clamp (pose 1)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Similar cross clamp (pose 2)</td>
<td>0.407</td>
<td>0.034</td>
<td>0.299</td>
<td>0.452</td>
</tr>
<tr>
<td>Small starter</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Large starter</td>
<td>0.505</td>
<td>0.055</td>
<td>0.398</td>
<td>0.581</td>
</tr>
<tr>
<td>Smaller cross clamp</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

[Holz et al. IROS 2015]
Different Lighting Conditions

- Artificial light and day light
- Only daylight
- Low light

- In all cases, the palette was successfully cleared.

[Holz et al. IROS 2015]
Deep Learning

[Schulz and Behnke, KI 2012]
GPU Implementations (CUDA)

- Affordable parallel computers
- General-purpose programming
- Convolutional
  - Local connectivity

[Scherer & Behnke, 2009] [Uetz & Behnke, 2009]
Image Categorization: NORB

- 10 categories, jittered-cluttered

- **Max-Pooling**, cross-entropy training

- Test error: 5.6% (LeNet7: 7.8%)

  [Scherer, Müller, Behnke, ICANN’10]
Image Categorization: LabelMe

- 50,000 color images (256x256)
- 12 classes + clutter (50%)

Error TRN: 3.77%; TST: 16.27%
Recall: 1,356 images/s

[Uetz, Behnke, ICIS2009]
Object-class Segmentation

- Class annotation per pixel
- Multi-scale input channels
- Evaluated on MSRC-9/21 and INRIA Graz-02 data sets
Object Detection in Images

- Bounding box annotation
- Structured loss that directly maximizes overlap of the prediction with ground truth bounding boxes
- Evaluated on two of the Pascal VOC 2007 classes

[Schulz, Behnke, ICANN 2014]
RGB-D Object-Class Segmentation

- Scale input according to depth
- Compute pixel height

![RGB-D Object-Class Segmentation Diagram](image)

<table>
<thead>
<tr>
<th>Method</th>
<th>floor</th>
<th>struct</th>
<th>furnit</th>
<th>prop</th>
<th>Class Avg.</th>
<th>Pixel Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>CW</td>
<td>84.6</td>
<td>70.3</td>
<td>58.7</td>
<td>52.9</td>
<td>66.6</td>
<td>65.4</td>
</tr>
<tr>
<td>CW+DN</td>
<td>87.7</td>
<td>70.8</td>
<td>57.0</td>
<td>53.6</td>
<td>67.3</td>
<td>65.5</td>
</tr>
<tr>
<td>CW+H</td>
<td>78.4</td>
<td>74.5</td>
<td>55.6</td>
<td>62.7</td>
<td>67.8</td>
<td>66.5</td>
</tr>
<tr>
<td>CW+DN+H</td>
<td>93.7</td>
<td>72.5</td>
<td>61.7</td>
<td>55.5</td>
<td>70.9</td>
<td>70.5</td>
</tr>
<tr>
<td>CW+DN+H+SP</td>
<td>91.8</td>
<td>74.1</td>
<td>59.4</td>
<td>63.4</td>
<td>72.2</td>
<td>71.9</td>
</tr>
<tr>
<td>CW+DN+H+CRF</td>
<td>93.5</td>
<td>80.2</td>
<td>66.4</td>
<td>54.9</td>
<td>73.7</td>
<td>73.4</td>
</tr>
<tr>
<td>Müller et al.[8]</td>
<td>94.9</td>
<td>78.9</td>
<td>71.1</td>
<td>42.7</td>
<td>71.9</td>
<td>72.3</td>
</tr>
<tr>
<td>Random Forest [8]</td>
<td>90.8</td>
<td>81.6</td>
<td>67.9</td>
<td>19.9</td>
<td>65.1</td>
<td>68.3</td>
</tr>
<tr>
<td>Couprie et al.[9]</td>
<td>87.3</td>
<td>86.1</td>
<td>45.3</td>
<td>35.5</td>
<td>63.6</td>
<td>64.5</td>
</tr>
<tr>
<td>Höft et al.[10]</td>
<td>77.9</td>
<td>65.4</td>
<td>55.9</td>
<td>49.9</td>
<td>62.3</td>
<td>62.0</td>
</tr>
<tr>
<td>Silberman [12]</td>
<td>68</td>
<td>59</td>
<td>70</td>
<td>42</td>
<td>59.7</td>
<td>58.6</td>
</tr>
</tbody>
</table>

NYU Depth V2

*CW is covering windows, H is height above ground, DN is depth normalized patch sizes. SP is averaged within superpixels and SVM-rewighted. CRF is a conditional random field over superpixels [8]. Structure class numbers are optimized for class accuracy.*

[Schulz, Höft, Behnke, ESANN 2015]
Neural Abstraction Pyramid

- Data-driven
- Analysis
- Feature extraction

- Model-driven
- Synthesis
- Feature expansion

Signals

- Grouping
- Competition
- Completion

Abstract features

[Behnke, LNCS 2766, 2003]
Iterative Interpretation

- Interpret most obvious parts first

- Use partial interpretation as context to resolve local ambiguities

[Behnke, LNCS 2766, 2003]
Local Recurrent Connectivity

Forward projection

Hyper column

Lateral projection

Backward projection

Feature map

Processor element

Layer

Less abstract

More abstract

[Behnke, LNCS 2766, 2003]
Neural Abstraction Pyramid for RGB-D Video Object-class Segmentation

- NYU Depth V2 contains RGB-D video sequences
- Recursive computation is efficient for temporal integration

<table>
<thead>
<tr>
<th>Method</th>
<th>Class Accuracies (%)</th>
<th>Average (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ground</td>
<td>struct</td>
</tr>
<tr>
<td>Höfft et al. [19]</td>
<td>77.9</td>
<td>65.4</td>
</tr>
<tr>
<td>Unidirectional + MS</td>
<td>73.4</td>
<td>66.8</td>
</tr>
<tr>
<td>Schulz et al. [20] (no height)</td>
<td>87.7</td>
<td>70.8</td>
</tr>
<tr>
<td>Unidirectional + SW</td>
<td><strong>90.0</strong></td>
<td><strong>76.3</strong></td>
</tr>
</tbody>
</table>

[Pavel, Schulz, Behnke, IJCNN 2015]
Geometric and Semantic Features for RGB-D Object-class Segmentation

- New **geometric** feature: distance from wall
- **Semantic** features pretrained from ImageNet
- Both help significantly

[Husain et al. under review]
Semantic Segmentation Priors for Object Discovery

- Combine bottom-up object discovery and semantic priors
- Semantic segmentation used to classify color and depth superpixels
- Higher recall, more precise object borders

[Garcia et al. under review]
RGB-D Object Recognition and Pose Estimation

- Use pretrained features from ImageNet

[Schwarz, Schulz, Behnke, ICRA2015]
Canonical View, Colorization

- Objects viewed from different elevation
- Render canonical view
- Colorization based on distance from center vertical

[Schwarz, Schulz, Behnke, ICRA2015]
Features Disentangle Data

- t-SNE embedding

- [Schwarz, Schulz, Behnke ICRA2015]
Recognition Accuracy

- Improved both category and instance recognition

<table>
<thead>
<tr>
<th>Method</th>
<th>Category Accuracy (%)</th>
<th>Instance Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RGB</td>
<td>RGB-D</td>
</tr>
<tr>
<td>Lai et al. [1]</td>
<td>74.3 ± 3.3</td>
<td>81.9 ± 2.8</td>
</tr>
<tr>
<td>Bo et al. [2]</td>
<td>82.4 ± 3.1</td>
<td>87.5 ± 2.9</td>
</tr>
<tr>
<td>PHOW[3]</td>
<td>80.2 ± 1.8</td>
<td>—</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>83.1 ± 2.0</strong></td>
<td><strong>88.3 ± 1.5</strong></td>
</tr>
<tr>
<td>Ours</td>
<td><strong>83.1 ± 2.0</strong></td>
<td><strong>89.4 ± 1.3</strong></td>
</tr>
</tbody>
</table>

Confusion

1: pitcher / coffe mug  
2: peach / sponge

[Schwarz, Schulz, Behnke, ICRA2015]
Conclusion

- Semantic perception in everyday environments is challenging
- Simple methods rely on strong assumptions (e.g. support plane)
- Depth helps with segmentation, allows for size normalization, geometric features, shape descriptors
- Deep learning methods work well
- Transfer of features from large data sets
- Many open problems, e.g. total scene understanding, incorporating physics, ...
Thanks for your attention!

Questions?