Learning Semantic Perception for Cognitive Robots

Sven Behnke
University of Bonn, Germany
Computer Science Institute VI
Autonomous Intelligent Systems
Some of Our Cognitive Robots

- Equipped with many sensors and DoFs
- Demonstration in complex scenarios

Soccer robot  Service robot  Exploration robot  Picking robot  MAV
Visual Perception of Soccer Scene

[Farazi & Behnke, RoboCup 2016]
RoboCup 2016 TeenSize Final
Robot Detection, Tracking & Identification

Robot Detection

HOG Feature Gathering

Gray

H

SVM

Projection to Egocentric

Heading Estimation

LSTM Network

Robot Identification

Annotated Image

Tracking

Wi-Fi

Robot Heading

[Farazi & Behnke, IROS 2017]
Robot Detection

• Based on HoG features

• Scan line feet estimation

[Farazi & Behnke, IROS 2017]
Visual Heading Estimation

• Dense HOG on upper half of detection
• SVM multiclass classifier
• 10 classes (36° each)

[Farazi & Behnke, IROS 2017]
Learning Data Association

- Recurrent neural network
- Training with simulated data

[Farazi & Behnke, IROS 2017]

- Fine-tuning on real data
Real-Robot Experiment

• Three Igus humanoid robots, observer in goal area
• Randomly chosen sequences, 3140 frames in total
• Partial, short term and long term occlusions, Single forward 4ms (∼250Hz)
RoboCup 2017 AdultSize Final
Cognitive Service Robot Cosero
Table-top Analysis and Grasp Planning

- Detection of clusters above horizontal plane
- Two grasps (top, side)
- Flexible grasping of many unknown objects

[Stückler et al, Robotics and Autonomous Systems, 2013]
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]
Learning and Tracking Object Models

• Modeling of objects by RGB-D-SLAM

• Real-time registration with current RGB-D frame
Deformable RGB-D-Registration

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

• Derived from the deformation field

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

[Stückler,Behnke,ICRA2014]
Tool use: Bottle Opener

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

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

- Simultaneous object modeling and motion segmentation

- Inference of a segment hierarchy

[Stückler, Behnke: IJCAI 2013]
Semantic Mapping

- Pixel-wise classification of RGB-D images by random forests
- Compare color / depth of regions
- Size normalization
- 3D fusion through RGB-D SLAM
- Evaluation on NYU depth v2

<table>
<thead>
<tr>
<th></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></td>
</tr>
</tbody>
</table>
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:

<table>
<thead>
<tr>
<th>Method</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>

Random forest
CRF prediction
Ground truth
Deep Learning

- Learning layered representations

[Schulz; Behnke, KI 2012]
Object-class Segmentation

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

[Schulz, Behnke, ESANN 2012]
Object Detection in Natural 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

- Covering windows segmented with CNN
- Scale input according to depth, compute pixel height

---

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

- Data-driven
- Analysis
- Feature extraction

Abstract features

- Model-driven
- Synthesis
- Feature expansion

Signals

- Grouping
- Competition
- Completion

[Behnke, Rojas, IJCNN 1998]
[Behnke, LNCS 2766, 2003]
Iterative Image Interpretation

- Interpret most obvious parts first
- Use partial interpretation as context to resolve local ambiguities
Neural Abstraction Pyramid for RGB-D Video Object-class Segmentation

- Recursive computation is efficient for temporal integration

[Image of Neural Abstraction Pyramid]

[Pavel, Schulz, Behnke, Neural Networks 2017]
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. RA-L 2016]
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. ICPR 2016]
RGB-D Object Recognition and Pose Estimation

[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]
Pretrained 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>83.1 ± 2.0</td>
<td>88.3 ± 1.5</td>
</tr>
<tr>
<td>Ours</td>
<td>83.1 ± 2.0</td>
<td>89.4 ± 1.3</td>
</tr>
</tbody>
</table>

- Confusion

1: pitcher / coffe mug
2: peach / sponge

[Schwarz, Schulz, Behnke, ICRA2015]
Autonomous Flight Near Obstacles

- Multimodal obstacle detection
  - 3D laser scanner
  - Stereo cameras
  - Ultrasound

[Droeschel et al.: Journal of Field Robotics, 2015]
Egocentric Laser-based 3D Mapping

• Motion compensation
  Distorted
  Undistorted

• Local multiresolution surfel maps
Allocentric 3D Map

- Registration of egocentric maps
- Global optimization of registration error by GraphSLAM

[Droeschel et al. JFR 2016]
Hierarchical Navigation

Mission plan

Allocentric planning

Egocentric planning

Obstacle avoidance

User

Operator station

Semantic map

Allocentric map

Onboard computer

Egocentric map

Obstacle map

Request

Mission planning

<0.02 Hz Observation poses

0.2 Hz Allocentr. plan

2 Hz Trajectory

20 Hz Speed

Copter
3D Simultaneous Localization and Mapping
Autonomous Flight in Warehouses

• Dual 3D laser scanner

• Omnidirectional cameras

• RFID reader
3D Map
Localization
Autonomous Mission in Warehouse
DJI Matrice 600 with Velodyne Puck
InventAIRy Final Demonstration

Fully Autonomous indoor flight without external tracking.
EuRoC Challenge 3: ChimneySpector
Mobile Manipulation Robot Momaro

- Four compliant legs ending in pairs of steerable wheels
- Anthropomorphically shaped upper body
- Sensor head
  - 3D laser scanner
  - IMU, cameras

[Schwarz et al. Journal of Field Robotics 2017]
Egress
Manipulation Operator Interface

- 3D head-mounted display
- 3D environment model + images
- 6D magnetic tracker

[Rodehutskors et al., Humanoids 2015]
Opening a Door
Local Multiresolution Surfel Map

- Registration and aggregation of 3D laser scans
- Local multi-resolution grid
- Surfel in grid cells

[Droeschel et al., Robotics and Autonomous Systems 2017]
Filtering Dynamic Objects

- Maintain occupancy in each cell
Allocentric 3D Mapping

- Registration of egocentric maps by graph optimization

[Droeschel et al., Robotics and Autonomous Systems 2017]
Operating a Switch
Debris Tasks
Team NimbRo Rescue

Best European Team (4th place overall), solved seven of eight tasks in 34 minutes
Stair Climbing

- Determine leg that most urgently needs to step
- Weight shift: sagittal, lateral, driving changes support
- Step to first possible foot hold after height change

[Schwarz et al., ICRA 2016]
DLR SpaceBot Cup 2015

• Mobile manipulation in rough terrain

[Schwarz et al., Frontiers on Robotics and AI 2016]
Autonomous Mission Execution

- 3D mapping, localization, mission and navigation planning
- 3D object perception and grasping

[Schwarz et al. Frontiers 2016]
Navigation Planning

• Costs from local height differences
• A* path planning

[Schwarz et al., Frontiers in Robotics and AI 2016]
3D Map
Considering Robot Footprint

- Costs for individual wheel pairs from height differences
- Base costs
- Non-linear combination yields 3D (x, y, θ) cost map

[Klamt and Behnke, IROS 2017]
3D Driving Planning (x, y, θ): A*

- 16 driving directions
- Orientation changes

=> Obstacle between wheels

[Klamt and Behnke, IROS 2017]
Making Steps

• If not drivable obstacle in front of a wheel
• Step landing must be drivable
• Support leg positions must be drivable

[Klamt and Behnke: IROS 2017]
Expanding Abstract Steps to Detailed Motion Sequences

[Klamt and Behnke: IROS 2017]
Planning for Challenging Scenarios

[Klamt and Behnke: IROS 2017]
New Sensor Head

- Continuously rotating Velodyne Puck VLP-16
  - 300,000 3D points/s
  - 100 m range
  - Spherical field of view
- Three wide-angle color cameras (total FoV $210 \times 103^\circ$)
- Kinect V2 RGB-D camera on pan-tilt unit
3D Map of Indoor+Outdoor Scene

[Droeschel et al., Robotics and Autonomous Systems 2017]

Sven Behnke: Learning Semantic Perception for Cognitive Robots
Navigation in allocentric laser map (colored points)
Amazon Picking Challenge

• Large variety of objects
• Unordered in shelf or tote
• Picking and stowing tasks

[Schwarz et al. ICRA 2017]
Deep Learning Semantic Segmentation

• Adapted from our segmentation of indoor scenes [Husain et al. RA-L 2016]

[Schwarz et al. ICRA 2017]
Combined Detection and Segmentation

[Schwarz et al. IJRR 2017]
Stowing
Picking
NimbRo Picking APC 2016 Results

- 2\textsuperscript{nd} Place Stowing (186 points)
- 3\textsuperscript{rd} Place Picking (97 points)

[Schwarz et al. IJRR 2017]
CENTAURO Workspace Perception Data Set

129 frames, 6 object classes

https://www.centauro-project.eu/data_multimedia/tools_data
Deep Learning Object Detection

[Johnson et al. 2015]
Detection of Tools

[Schwarz et al. IJRR 2017]
Semantic Segmentation

- Deep CNN

Pixel-wise accuracy:

<table>
<thead>
<tr>
<th></th>
<th>Clamp</th>
<th>Door handle</th>
<th>Driller</th>
<th>Extension</th>
<th>Stapler</th>
<th>Wrench</th>
<th>Background</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.727</td>
<td>0.751</td>
<td>0.769</td>
<td>0.889</td>
<td>0.775</td>
<td>0.734</td>
<td>0.992</td>
<td>0.805</td>
</tr>
</tbody>
</table>

[Husain et al. RA-L 2016]
MBZIRC Challenge 2
Wrench Selection: Detection of Tool Ends
Picking Copter DJI Matrice 100

- Wide-angle downward looking color camera
- Electromagnetic gripper
- Laser-distance sensor to ground
- Dual-core PC
MBZIRC Team NimbRo
EuRoC C1 Robolink Feeder: Bin Picking
Part Pose Estimation

- Two convolutional neural networks

- Training with synthetic depth images

### PoseNet

### SymNet

[Koo et al. CASE 2017]
Robolink Feeder: Regrasping and Placing
Amazon Robotics Challenge 2017

- Quick learning of novel objects
- Training with rendered scenes
RefineNet

• Increase resolution by using features from the higher resolution
• Coarse-to-fine semantic segmentation

[Lin et al. CVPR 2017]
Object Capture and Scene Rendering

- Turn table + DLSR

Rendered scenes
ARC 2017 Perception Example
NimbRo Picking 2017 Team

- 2\textsuperscript{nd} place Pick
- 2\textsuperscript{nd} place Stow-and-Pick Final
Object Pose Estimation

- Use upper layer of RefineNet as input
- Predict pose coordinates for one segment
Conclusions

• Semantic perception is challenging
• Simple methods rely on strong assumptions
• Depth helps with segmentation, allows for size normalization, geometric features, shape descriptors
• Deep learning methods work well
• Transfer of features from large data sets
• Synthetic training
• Many open problems, e.g. total scene understanding, incorporating physics, ...
Questions?