(Deep) Learning for Robot Perception and Navigation

Wolfram Burgard
Deep Learning for Robot Perception (and Navigation)

Lifeng Bo, Claas Bollen, Thomas Brox, Andreas Eitel, Dieter Fox, Gabriel L. Oliveira, Luciano Spinello, Jost Tobias Springenberg, Martin Riedmiller, Michael Ruhnke, Abhinav Valada
Perception in Robotics

- Robot perception is a challenging problem and involves many different aspects such as
  - Scene understanding
  - Object detection
  - Detection of humans

- Goal: improve perception in robotics scenarios using state-of-the-art deep learning methods
Why Deep Learning?

- Multiple layers of abstraction provide an advantage for solving complex pattern recognition problems
- Successful in computer vision for detection, recognition, and segmentation problems
- One set of techniques can serve different fields and be applied to solve a wide range of problems
What Our Robots Should Do

- RGB-D object recognition
- Images human part segmentation
- Sound terrain classification

- Asphalt
- Mowed Grass
- Grass
Multimodal Deep Learning for Robust RGB-D Object Recognition

Andreas Eitel, Jost Tobias Springenberg, Martin Riedmiller, Wolfram Burgard

[IROS 2015]
RGB-D Object Recognition
RGB-Depth Object Recognition

- Learned features + classifier
  - Sparse coding networks [Bo et. al 2012]
  - Deep CNN features [Schwarz et. al 2015]

- End-to-end learning / Deep learning
  - Convolutional recurrent neural networks [Socher et. al 2012]
Often too little Data for Deep Learning Solutions

Deep networks are hard to train and require large amounts of data

- Lack of large amount of labeled training data for RGB-D domain
- How to deal with limited sizes of available datasets?
Data often too Clean for Deep Learning Solutions

Large portion of RGB-D data is recorded under controlled settings

- How to improve recognition in real-world scenes when the training data is “clean”?
- How to deal with sensor noise from RGB-D sensors?
Solution: Transfer Deep RGB Features to Depth Domain

Both domains share similar features such as edges, corners, curves, ...
Solution: Transfer Deep RGB Features to Depth Domain

Depth domain

Pre-trained RGB CNN

RGB domain

Transfer*

Depth encoding

Fine-tune

Re-train network features for depth

* Similar to [Schwarz et. al 2015, Gupta et. al 2014]
Solution: Transfer Deep RGB Features to Depth Domain

* Similar to [Schwarz et. al 2015, Gupta et. al 2014]
Multimodal Deep Convolutional Neural Network

- Two input modalities
- Late fusion network
- 10 convolutional layers
- Max pooling layers
- 4 fully connected layers
- Softmax classifier

2xAlexNet + fusion net
How to Encode Depth Images?

- Distribute depth over color channels
  - Compute min and max value of depth map
  - Shift depth map to min/max range
  - Normalize depth values to lie between 0 and 255
  - Colorize image using jet colormap (red = near, blue = far)
- Depth encoding improves recognition accuracy by 1.8 percentage points

![RGB](image1) ![Raw depth](image2) ![Colorized depth](image3)
Solution: Noise-aware Depth Feature Learning

“Clean” training data → Noise samples → Noise adaptation → Classify
Training with Noise Samples

- Randomly sample noise for each training batch
- Shuffle noise samples

Input image

Training batch

Noise samples: 50,000
RGB Network Training

- Maximum likelihood learning
- Fine-tune from pre-trained AlexNet weights
Depth Network Training

- Maximum likelihood learning
- Fine-tune from pre-trained AlexNet weights

\[ p(y \mid d) \]
Fusion Network Training

- Fusion layers automatically learn to combine feature responses of the two network streams
- During training, weights in first layers stay fixed
UW RGB-D Object Dataset

Category-Level Recognition [%] (51 categories)

<table>
<thead>
<tr>
<th>Method</th>
<th>RGB</th>
<th>Depth</th>
<th>RGB-D</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN-RNN</td>
<td>80.8</td>
<td>78.9</td>
<td>86.8</td>
</tr>
<tr>
<td>HMP</td>
<td>82.4</td>
<td>81.2</td>
<td>87.5</td>
</tr>
<tr>
<td>CaRFs</td>
<td>N/A</td>
<td>N/A</td>
<td>88.1</td>
</tr>
<tr>
<td>CNN Features</td>
<td>83.1</td>
<td>N/A</td>
<td>89.4</td>
</tr>
</tbody>
</table>

[Lai et al, 2011]
UW RGB-D Object Dataset

Category-Level Recognition [%] (51 categories)

<table>
<thead>
<tr>
<th>Method</th>
<th>RGB</th>
<th>Depth</th>
<th>RGB-D</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN-RNN</td>
<td>80.8</td>
<td>78.9</td>
<td>86.8</td>
</tr>
<tr>
<td>HMP</td>
<td>82.4</td>
<td>81.2</td>
<td>87.5</td>
</tr>
<tr>
<td>CaRFs</td>
<td>N/A</td>
<td>N/A</td>
<td>88.1</td>
</tr>
<tr>
<td>CNN Features</td>
<td>83.1</td>
<td>N/A</td>
<td>89.4</td>
</tr>
<tr>
<td>This work, Fus-CNN</td>
<td><strong>84.1</strong></td>
<td><strong>83.8</strong></td>
<td><strong>91.3</strong></td>
</tr>
</tbody>
</table>

[Lai et. al, 2011]
Confusion Matrix

Label | Prediction
--- | ---
mushroom | garlic
pitcher | coffee mug
peach | garlic
Recognition in Noisy RGB-D Scenes

Recognition using annotated bounding boxes

Noise adapt. = correct prediction
No adapt. = false prediction

Category-Level Recognition [%] depth modality (6 categories)

<table>
<thead>
<tr>
<th>Noise adapt.</th>
<th>flash-light</th>
<th>cap</th>
<th>bowl</th>
<th>soda can</th>
<th>cereal box</th>
<th>coffee mug</th>
<th>class avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>97.5</td>
<td>68.5</td>
<td>66.5</td>
<td>66.6</td>
<td>96.2</td>
<td>79.1</td>
<td>79.1</td>
</tr>
<tr>
<td>√</td>
<td>96.4</td>
<td>77.5</td>
<td>69.8</td>
<td>71.8</td>
<td>97.6</td>
<td>79.8</td>
<td>82.1</td>
</tr>
</tbody>
</table>
Deep Learning for RGB-D Object Recognition

- Novel RGB-D object recognition for robotics
- Two-stream CNN with late fusion architecture
- Depth image transfer and noise augmentation training strategy
- State of the art on UW RGB-D Object dataset for category recognition: 91.3%
- Recognition accuracy of 82.1% on the RGB-D Scenes dataset
Deep Learning for Human Part Discovery in Images

Gabriel L. Oliveira, Abhinav Valada, Claas Bollen, Wolfram Burgard, Thomas Brox

[submitted to ICRA 2016]
Deep Learning for Human Part Discovery in Images

- Human-robot interaction

- Robot rescue
Deep Learning for Human Part Discovery in Images

- Dense prediction can provide pixel classification of the image
- Human part segmentation is naturally challenging due to
  - Non-rigid aspect of body
  - Occlusions

PASCAL Parts  MS COCO  Freiburg Sitting
Network Architecture

- Fully convolutional network
  - Contraction and expansion of network input
  - Up-convolution operation for expansion
- Pixel input, pixel output
Experiments

- Evaluation of approach on
  - Publicly available computer vision datasets
  - Real-world datasets with ground and aerial robots
- Comparison against state-of-the-art semantic segmentation approach: FCN proposed by Long et al. [1]

Data Augmentation

Due to the low number of images in the available datasets, augmentation is crucial

- Spatial augmentation (rotation + scaling)

- Color augmentation
# PASCAL Parts Dataset

- **PASCAL Parts, 4 classes, IOU**

<table>
<thead>
<tr>
<th>Method</th>
<th>Head</th>
<th>Torso</th>
<th>Arms</th>
<th>Legs</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCN</td>
<td>70.74</td>
<td>60.62</td>
<td>48.44</td>
<td>50.38</td>
<td>57.35</td>
</tr>
<tr>
<td>Ours</td>
<td>75.08</td>
<td>64.81</td>
<td>55.61</td>
<td>56.72</td>
<td>63.03</td>
</tr>
<tr>
<td>Ours (Spatial)</td>
<td>80.49</td>
<td>74.39</td>
<td>67.17</td>
<td>70.39</td>
<td>73.00</td>
</tr>
<tr>
<td>Ours (Spatial + Color)</td>
<td><strong>83.24</strong></td>
<td><strong>79.41</strong></td>
<td><strong>73.73</strong></td>
<td><strong>76.52</strong></td>
<td><strong>78.23</strong></td>
</tr>
</tbody>
</table>

- **PASCAL Parts, 14 classes, IOU**

<table>
<thead>
<tr>
<th>Method</th>
<th>Head</th>
<th>Torso</th>
<th>L U arm</th>
<th>L LW arm</th>
<th>L hand</th>
<th>R U hand</th>
<th>R LW arm</th>
<th>R hand</th>
<th>R U leg</th>
<th>R LW leg</th>
<th>R foot</th>
<th>L U leg</th>
<th>L LW leg</th>
<th>L foot</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCN</td>
<td>74.0</td>
<td>66.2</td>
<td>56.6</td>
<td>46.0</td>
<td>34.1</td>
<td>58.9</td>
<td>44.1</td>
<td>31.0</td>
<td>49.3</td>
<td>44.5</td>
<td>40.8</td>
<td>48.5</td>
<td>47.6</td>
<td>41.2</td>
<td>53.1</td>
</tr>
<tr>
<td>Ours (Spatial)</td>
<td>81.8</td>
<td>78.0</td>
<td>69.5</td>
<td>63.1</td>
<td>59.0</td>
<td>71.2</td>
<td>63.0</td>
<td>58.7</td>
<td>65.4</td>
<td>60.6</td>
<td>52.0</td>
<td>67.9</td>
<td>60.3</td>
<td>50.0</td>
<td>66.9</td>
</tr>
<tr>
<td>Ours (Spatial + Color)</td>
<td><strong>84.0</strong></td>
<td><strong>81.5</strong></td>
<td><strong>74.1</strong></td>
<td><strong>68.0</strong></td>
<td><strong>64.0</strong></td>
<td><strong>75.4</strong></td>
<td><strong>67.4</strong></td>
<td><strong>61.9</strong></td>
<td><strong>72.4</strong></td>
<td><strong>67.1</strong></td>
<td><strong>56.9</strong></td>
<td><strong>73.0</strong></td>
<td><strong>66.1</strong></td>
<td><strong>57.7</strong></td>
<td><strong>71.7</strong></td>
</tr>
</tbody>
</table>

*Note: R = right, L = left, U = upper, LW = lower.*
Freiburg Sitting People Part Segmentation Dataset

- We present a novel dataset for human part segmentation in wheelchairs

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
<th>IOU</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCN</td>
<td>59.69</td>
<td>43.17</td>
</tr>
<tr>
<td>Ours (Trained on PASCAL)</td>
<td>78.04</td>
<td>59.84</td>
</tr>
<tr>
<td>Ours (2 people train - 4 people test)</td>
<td>81.78</td>
<td>64.10</td>
</tr>
</tbody>
</table>
Robot Experiments

- Range experiments with ground robot
- Aerial platform for disaster scenario
  - Segmentation under severe body occlusions
Range Experiments

Recorded using Bumblebee camera

- Robust to radial distortion
- Robust to scale

(a) 1.0 meter  (b) 2.0 meters
(c) 3.0 meters  (d) 4.0 meters
(e) 5.0 meters  (f) 6.0 meters

![Graph showing Mean IOU (%) vs Distance (m)]
Freiburg People in Disaster

Dataset designed to test severe occlusions

<table>
<thead>
<tr>
<th>Method</th>
<th>Head</th>
<th>Torso</th>
<th>Arms</th>
<th>Legs</th>
<th>IOU</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCN</td>
<td>52.71</td>
<td>62.49</td>
<td>35.04</td>
<td>43.25</td>
<td>43.20</td>
</tr>
<tr>
<td>Ours</td>
<td>80.56</td>
<td>79.45</td>
<td>63.93</td>
<td>64.91</td>
<td>71.99</td>
</tr>
</tbody>
</table>
Future Work

- Investigate the potential for human keypoint annotation
- Real-time part segmentation for small hardware
- Human part segmentation in videos
Deep Feature Learning for Acoustics-based Terrain Classification

Abhinav Valada, Luciano Spinello, Wolfram Bugard

[ISRR 2015]
Motivation

Robots are increasingly being used in unstructured real-world environments
Motivation

Optical sensors are highly sensitive to visual changes.
Motivation

Use sound from vehicle-terrain interactions to classify terrain
Network Architecture

- Novel architecture designed for unstructured sound data
- Global pooling gathers statistics of learned features across time
Data Collection

Wood  Linoleum  Carpet  P3-DX

Asphalt  Mowed Grass  Grass  Paving  Cobble Stone  Offroad
# Results - Baseline Comparison

(300ms window)

<table>
<thead>
<tr>
<th>Features</th>
<th>SVM Linear</th>
<th>SVM RBF</th>
<th>k-NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ginna [1]</td>
<td>44.87 ± 0.70</td>
<td>37.51 ± 0.74</td>
<td>57.26 ± 0.60</td>
</tr>
<tr>
<td>Spectral [2]</td>
<td>84.48 ± 0.36</td>
<td>78.65 ± 0.45</td>
<td>76.02 ± 0.43</td>
</tr>
<tr>
<td>Ginna &amp; Shape [3]</td>
<td>85.50 ± 0.34</td>
<td>80.37 ± 0.55</td>
<td>78.17 ± 0.37</td>
</tr>
<tr>
<td>MFCC &amp; Chroma [4]</td>
<td>88.95 ± 0.21</td>
<td><strong>88.55 ± 0.20</strong></td>
<td>88.43 ± 0.15</td>
</tr>
<tr>
<td>Trimbral [5]</td>
<td>89.07 ± 0.12</td>
<td>86.74 ± 0.25</td>
<td>84.82 ± 0.54</td>
</tr>
<tr>
<td>Cepstral [6]</td>
<td><strong>89.93 ± 0.21</strong></td>
<td>78.93 ± 0.62</td>
<td><strong>88.63 ± 0.06</strong></td>
</tr>
</tbody>
</table>

86.92% improvement over the previous state of the art

---

Robustness to Noise

Per-class Precision
Noise Adaptive Fine-Tuning

Avg. accuracy of 99.57% on the base model
Real-World Stress Testing

Avg. accuracy of 98.54%

- True Positives
- False Positives
Can you guess the terrain?

Social Experiment

- Avg. human performance = 24.66%
- Avg. network performance = 99.5%
- Go to deepterrain.cs.uni-freiburg.de
- Listen to five sound clips of a robot traversing on different terrains
- Guess what terrain they are
Conclusions

- Classifies terrain using only sound
- State-of-the art performance in proprioceptive terrain classification
- New DCNN architecture outperforms traditional approaches
- Noise adaptation boosts performance
- Experiments with a low-quality microphone demonstrates robustness
Overall Conclusions

- Deep networks are a promising approach to solve complex perception problems in robotics
- The key challenges are
  - finding the proper architecture
  - using proper data augmentation strategies