Object Recognition and Segmentation in Indoor Scenes from RGB-D Images

Md. Alimoor Reza and Jana Kosecka
George Mason University
Motivation

- Semantic understanding of the open-indoor scenes paves the way for high-level robotics tasks.
- The goal of semantic segmentation is to label all the pixels in the image belonging to predefined number of semantic object categories.
- Many tasks do not require association of semantic label with each pixel, but only labels needed for the task (navigation: road/floor, localization: landmarks/ vegetation/structures; generic objects) [Cadena-Kosecka, 2013]
- Here we describe the work to obtain finer categorization of object classes (e.g., **table**, **bed**, **lab**, **sink**)
Existing Approaches

- Generate high-quality bottom up segmentation from low-level grouping cues
- Rich feature computation
- Region Classification
  - Gupta et al. '2013, Ren et al. '2012

- Simple superpixels and feature computation.
- Inference in CRF for multi-class semantic segmentation.
  - Silberman'2011, Cadena'2013

- Generate multiple object proposals (using CPMC)

- Augment second order statistics to local descriptors.
  - Carreira et al. '2012

- Holistic Scene understanding
  - Utrasun'2013

- Stacked hierarchical labeling: CRF inference over hierarchical regions
  - Munoz, 2012
Our Approach

- Task constrained application may seek a single object of interest.
- Easily extended to multiple object categories.
- We formulate the problem of recognition and segmentation of objects in indoor scenes as a binary **object-of-interest vs background** segmentation task.

**Our choices:**
- Regular sized regions from efficient low-level superpixel segmentation.
- Rich features.
- Efficient inference in CRF.
Conditional Random Field (CRF)

- Learn per-category object grouping in a CRF framework

\[
p(y|z) = \frac{1}{Z(z)} \exp\left( w_1 \sum_{i}^{V} \theta_d(y_i, z) + w_2 \sum_{(i,j)}^{E} \theta_{pc}(y_i, y_j, z) + w_3 \sum_{(i,j)}^{E} \theta_{px}(y_i, y_j, z) \right)
\]

- **Unary:**
  - Computed from the probabilistic output of the AdaBoost classifier.

\[
\theta_d(y_i, z) = -\log(P_i(y_i|z))
\]

- **Pairwise:**
  - Color (Lab color space) contrast of super-pixels and their label difference.
  - Spatial contrast of super-pixels and their label difference.
Features

- The set of observation in our local prior $P_i(y_i|z)$ are computed for each SLIC superpixel in image.

- **Color**: 75 bin histogram of color in HSV color space.
- **Texture**: 240 bin histogram of texture by convolving image with oriented filters.
Features

- **Geometric**: geometric features capturing local and global geometry [Cadena et al. 2013].

- **Generic**: a set of generic features adopted from Gupta et al. 2013.

- Total dimension **386**.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Color</td>
<td>75</td>
</tr>
<tr>
<td>Texture</td>
<td>240</td>
</tr>
<tr>
<td>Geometric</td>
<td>11</td>
</tr>
<tr>
<td>Generic</td>
<td>60</td>
</tr>
</tbody>
</table>
Classifier with Negative Mining

- Learned one-vs-all AdaBoost classifiers for each object.
- Maintained almost an equal proportion of positive-negative sample.
- Negative mining to select samples from other objects that co-occur with the object of interest.

Co-occurrence Matrix  

Histogram of co-occurrence for book object.  

A scene with books on the bookshelf.
System Overview

Feature computation
- Color, Texture, Geometric, Generic

Object-specific classifier (AdaBoost) with negative mining

Object recognition and segmentation

CRF learning and inference
Results

- NYUD-V2 dataset.
- 1449 images (benchmark split 795 training and 654 test images).
- Selected 34 most frequent objects (Bed, Chair, Table, Light etc) of furniture and prop. categories.
- Evaluation:
  - Evaluated on the test images, where the object is present
  - Jaccard Index \[ JI = \frac{|P \cap G|}{|P \cup G|} \]
  - Per-class accuracy
### Comparison in Jaccard Index

<table>
<thead>
<tr>
<th>Approach</th>
<th>Bed</th>
<th>Sofa</th>
<th>Chair</th>
<th>Table</th>
<th>Window</th>
<th>Bookshelf</th>
<th>TV</th>
<th>Bag</th>
<th>Bathtub</th>
<th>Blinds</th>
<th>Books</th>
<th>Box</th>
<th>Cabinet</th>
<th>Clothes</th>
<th>Counter</th>
<th>Curtain</th>
<th>Desks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Silberman [5]</td>
<td>40.00</td>
<td>25.00</td>
<td>32.00</td>
<td>21.00</td>
<td>30.00</td>
<td>23.00</td>
<td>5.70</td>
<td>0.00</td>
<td>0.00</td>
<td>40.00</td>
<td>5.50</td>
<td>0.13</td>
<td>33.00</td>
<td>6.50</td>
<td>33.00</td>
<td>27.00</td>
<td>4.60</td>
</tr>
<tr>
<td>Ren [4]</td>
<td>42.00</td>
<td>28.00</td>
<td>33.00</td>
<td>17.00</td>
<td>28.00</td>
<td>17.00</td>
<td>19.00</td>
<td>1.20</td>
<td>7.80</td>
<td>27.00</td>
<td>15.00</td>
<td>3.30</td>
<td>37.00</td>
<td>9.50</td>
<td>39.00</td>
<td>28.00</td>
<td>10.00</td>
</tr>
<tr>
<td>Gupta [2]</td>
<td>55.00</td>
<td>44.00</td>
<td>40.00</td>
<td>30.00</td>
<td>33.00</td>
<td>20.00</td>
<td>9.30</td>
<td>0.65</td>
<td>33.00</td>
<td>44.00</td>
<td>4.40</td>
<td>4.80</td>
<td>48.00</td>
<td>6.90</td>
<td>47.00</td>
<td>34.00</td>
<td>10.00</td>
</tr>
<tr>
<td>Ours (unary)</td>
<td>50.64</td>
<td>37.44</td>
<td>25.00</td>
<td>19.19</td>
<td>25.93</td>
<td>23.88</td>
<td>26.40</td>
<td>3.28</td>
<td>32.12</td>
<td>29.77</td>
<td>9.17</td>
<td>2.89</td>
<td>27.42</td>
<td>9.79</td>
<td>34.68</td>
<td>25.59</td>
<td>21.04</td>
</tr>
<tr>
<td>Ours (CRF)</td>
<td>56.85</td>
<td>42.29</td>
<td>31.44</td>
<td>20.78</td>
<td>30.16</td>
<td>30.29</td>
<td>34.97</td>
<td>3.00</td>
<td>32.95</td>
<td>33.09</td>
<td>10.06</td>
<td>3.99</td>
<td>29.34</td>
<td>10.04</td>
<td>33.82</td>
<td>30.11</td>
<td>23.35</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Approach</th>
<th>Door</th>
<th>Dresser</th>
<th>Floor-mat</th>
<th>Lamp</th>
<th>Mirror</th>
<th>Night-stand</th>
<th>Paper</th>
<th>Person</th>
<th>Picture</th>
<th>Pillow</th>
<th>Refrigerator</th>
<th>Shelves</th>
<th>Shower-curtain</th>
<th>Sink</th>
<th>Toilet</th>
<th>Towel</th>
<th>Whiteboard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Silberman [5]</td>
<td>5.90</td>
<td>13.00</td>
<td>7.20</td>
<td>16.00</td>
<td>4.40</td>
<td>13.00</td>
<td>6.60</td>
<td>36.00</td>
<td>19.00</td>
<td>3.60</td>
<td>3.30</td>
<td>3.60</td>
<td>25.00</td>
<td>27.00</td>
<td>0.11</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Ren [4]</td>
<td>13.00</td>
<td>7.00</td>
<td>20.00</td>
<td>14.00</td>
<td>18.00</td>
<td>12.00</td>
<td>14.00</td>
<td>32.00</td>
<td>20.00</td>
<td>1.90</td>
<td>6.10</td>
<td>5.40</td>
<td>29.00</td>
<td>35.00</td>
<td>13.00</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td>Gupta [2]</td>
<td>8.30</td>
<td>22.00</td>
<td>22.00</td>
<td>6.80</td>
<td>19.00</td>
<td>20.00</td>
<td>16.00</td>
<td>40.00</td>
<td>28.00</td>
<td>15.00</td>
<td>5.10</td>
<td>18.00</td>
<td>26.00</td>
<td>50.00</td>
<td>14.00</td>
<td>37.00</td>
<td>36.25</td>
</tr>
<tr>
<td>Ours (unary)</td>
<td>14.72</td>
<td>32.35</td>
<td>32.81</td>
<td>6.68</td>
<td>23.09</td>
<td>16.22</td>
<td>7.64</td>
<td>19.54</td>
<td>17.93</td>
<td>16.16</td>
<td>16.86</td>
<td>10.67</td>
<td>25.54</td>
<td>10.98</td>
<td>26.06</td>
<td>7.62</td>
<td>36.25</td>
</tr>
<tr>
<td>Ours (CRF)</td>
<td>17.16</td>
<td>35.73</td>
<td>34.19</td>
<td>12.14</td>
<td>27.41</td>
<td>21.54</td>
<td>10.07</td>
<td>30.31</td>
<td>22.21</td>
<td>22.98</td>
<td>20.59</td>
<td>13.46</td>
<td>26.84</td>
<td>11.04</td>
<td>38.65</td>
<td>8.61</td>
<td>37.69</td>
</tr>
</tbody>
</table>

### References


4. X. Ren, L. Bo, and D. Fox. RGB-(D) scene labeling: Features and algorithms. (CVPR), 2012.


<table>
<thead>
<tr>
<th>Approach</th>
<th>Mean JI</th>
</tr>
</thead>
<tbody>
<tr>
<td>[2]</td>
<td>23.92</td>
</tr>
<tr>
<td>Ours</td>
<td>24.92</td>
</tr>
</tbody>
</table>

Table: Summary of results in Jaccard Index metric.
## Comparison in Per-class Accuracy

<table>
<thead>
<tr>
<th></th>
<th>Bed</th>
<th>Sofa</th>
<th>Chair</th>
<th>Table</th>
<th>Window</th>
<th>Books</th>
<th>TV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Couprie $[1]$</td>
<td>38.4</td>
<td>24.6</td>
<td>34.1</td>
<td>10.2</td>
<td>15.9</td>
<td>13.7</td>
<td>6.0</td>
</tr>
<tr>
<td>Hermans $[3]$</td>
<td>68.4</td>
<td>28.5</td>
<td>41.9</td>
<td>27.1</td>
<td>46.1</td>
<td>45.4</td>
<td>38.4</td>
</tr>
<tr>
<td>Ours</td>
<td>87.8</td>
<td>86.1</td>
<td>82.7</td>
<td>78.0</td>
<td>78.1</td>
<td>71.7</td>
<td>81.8</td>
</tr>
</tbody>
</table>

### References
Results

RGB  Ground truth  Without CRF  CRF
Results

RGB  Ground truth  Without CRF  CRF
Conclusion & Future Work

- Task constrained formulation yields simple binary CRF (learning and inference can be efficient).
- Category level supervision at the level of learning super-pixel grouping rules in CRF setting.
- Strong informative for finer discrimination
- Contextual relationships of object co-occurrence are effective in hard negative mining.

- Combine binary segmentations.
- Explore anytime approaches to tackle features computation and efficiency issues.