3D Object Detection

Zhen Li
CSC 2541 Presentation
Mar 8th, 2016
Object Detection: 2D vs 3D

Video (Chen et al. 2015)

3D Object Proposals for Accurate Object Class Detection

NIPS 2015

Xiaozhi Chen1, Kaustav Kunku2, Yukun Zhu2, Andrew Bereshaw2
Huimin Ma1, Sanja Fidler2, Raquel Urtasun2

1Tsinghua University 2University of Toronto
3D Object Detection: Motivation

• 2D bounding boxes are not sufficient
  • Lack of 3D pose, Occlusion information, and 3D location

(Figure from Felzenszwalb et al. 2010)  (Figure from Xiang et al. 2015)
3D Object Detection: Challenge

- Occlusion/Truncation: Only a small portion of the surface is visible
  - Leader board from KITTI website

<table>
<thead>
<tr>
<th></th>
<th>Easy</th>
<th>Moderate</th>
<th>Hard</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SubCNN</td>
<td>90.49%</td>
<td>87.88%</td>
</tr>
<tr>
<td>2</td>
<td>DJML</td>
<td>90.67%</td>
<td>87.51%</td>
</tr>
<tr>
<td>3</td>
<td>3DOP</td>
<td>91.44%</td>
<td>86.10%</td>
</tr>
<tr>
<td>4</td>
<td>Mono3D</td>
<td>88.31%</td>
<td>85.66%</td>
</tr>
<tr>
<td>5</td>
<td>3DVP</td>
<td>86.92%</td>
<td>74.59%</td>
</tr>
</tbody>
</table>

Easy: Max. occlusion 15%
Moderate: Max. occlusion 30%
Hard: Max. occlusion 50%
Outline

• Overview with contributions

• Main motivation

• Technical approach

• Experimental evaluation

• Discussion
Data-Driven 3D Voxel Patterns for Object Category Recognition

Yu Xiang\textsuperscript{1,2}, Wongun Choi\textsuperscript{3}, Yuanqing Lin\textsuperscript{3}, and Silvio Savarese\textsuperscript{1}
\textsuperscript{1}Stanford University, \textsuperscript{2}University of Michigan at Ann Arbor, \textsuperscript{3}NEC Laboratories America, Inc.
yuxiang@umich.edu, \{wongun, ylin\}@nec-labs.com, ssilvio@stanford.edu
High-level Overview

• Propose a novel object representation: 3D Voxel Pattern (3DVP)
  • Appearance, 3D shape, and occlusion masks

(Figure from Xiang et al. 2015)
High-level Overview

• Propose a novel object representation: 3D Voxel Pattern (3DVP)
  • Appearance, 3D shape, and occlusion masks

• Train specialized 3DVP detectors which are capable of:
  • 2D Object detection
  • Segmentation mask, occlusion or truncation boundaries
  • 3D localization, 3D pose

(Figure from Xiang et al. 2015)
High-level Overview

• Propose a novel object representation: 3D Voxel Pattern (3DVP)
  • Appearance, 3D shape, and occlusion masks
• Train specialized 3DVP detectors which are capable of:
  • 2D Object detection
  • Segmentation mask, occlusion or truncation boundaries
  • 3D localization, 3D pose
• Experiments on the KITTI benchmark and the OutdoorScene dataset
  • Improve the state-of-the-art results on detection and pose estimation with notable margins (6% in difficult level of KITTI)
Motivations

• What are the key challenges in this topic?
  • Occlusion/Truncation
    • Train partial object detectors for visible parts of objects (Wu and Nevatia 2005; Wojek et al. 2011; Xiang and Savarese 2013)

(Figure from Xiang et al. 2015)
Motivations

• What are the key challenges in this topic?
  • Occlusion/Truncation
  • Shape variation: Intra-class changes should be modeled
    • Discover and learn object sub-categories

(Figure from Xiang et al. 2015)
Motivations

• What are the key challenges in this topic?
  • Occlusion/Truncation
  • Shape variation: Intra-class changes should be modeled
  • Viewpoint: Multiview object detection in 3D
    • Built from various 2D images (Yan et al. 2007; Glasner et al. 2011)
    • Constructed using CAD models (Liebelt et al. 2008)

(Figure from Xiang et al. 2015)
Technical approach

• Training: Generate 3D Voxel Exemplars
  • A triplet of 2D image of the object, its 2D segmentation, and its 3D voxel model

(Figures from Xiang et al. 2015)
Technical approach

- Training: Generate 3D Voxel Exemplars
  - 3D CAD model association and registration
  - Project 3D CAD models to the image
  - Label 2D segmentation mask and 3D voxel model
  - Generate a 3D voxel exemplar

(Figures from Xiang et al. 2015)
Technical approach

• Training: Generate *3D Voxel Exemplars*
  • A triplet of 2D image of the object, its 2D segmentation, and its 3D voxel model

(Figures from Xiang et al. 2015)
Technical approach

- **Training**: Build a representative set of 3DVPs

(Figures from Xiang et al. 2015)
Technical approach

• Training: Build a representative set of 3DVPs

Figure 5. Examples of 3D clusters from the KITTI dataset.

(Figures from Xiang et al. 2015)
Technical approach

- **Training:** Build a representative set of 3DVPs
  - Define the 3D voxel exemplar feature vector $\mathbf{x}$ with dimension $N^3$
    - Encoding: 0 for empty voxels, 1 for visible voxels, 2 for self-occluded voxels, 3 for voxels occluded by other objects, and 4 for truncated voxels.
  - Define the similarity metric:
    \[
    s(\mathbf{x}_1, \mathbf{x}_2) = \frac{|S|}{N^3} \sum_{i=1}^{N^3} \mathbb{1}(x_1^i = x_2^i) \cdot w(x_1^i),
    \]
    \[
    \text{s.t., } \sum_{i=0}^{|S|-1} w(i) = 1,
    \]
Technical approach

- **Training**: Build a representative set of 3DVPs
  - Define the 3D voxel exemplar feature vector $x$ with dimension $N^3$
  - Encoding: 0 for empty voxels, 1 for visible voxels, 2 for self-occluded voxels, 3 for voxels occluded by other objects, and 4 for truncated voxels.
  - Define the similarity metric:
    $$s(x_1, x_2) = \frac{|S|}{N^3} \sum_{i=1}^{N^3} I(x_1^i) = 1$$
    s.t., $\sum_{i=0}^{|S|-1} w(i) = |
  - **Employ clustering algorithms**
    - K-means
    - **Affinity Propagation (AP) (Frey and Dueck 2007)**

(Video from http://www.psi.toronto.edu/affinitypropagation/)
Technical approach

- Training: Train 3DVP Detectors
  - SVM-based detectors for KITTI (Malisiewicz et al. 2011)
  - Boosting detector for KITTI
    - Aggregated Channel Features (ACF) (Dollár et al. 2014)

(Figures from Xiang et al. 2015)
Technical approach

• Training: Train 3DVP Detectors
  • SVM-based detectors for KITTI (Malisiewicz et al. 2011)
  • Boosting detector for KITTI
    • Aggregated Channel Features (ACF) (Dollár et al. 2014)

![Diagram of ACF detector](Images from Dollár et al. 2014)

Fig. 8. Overview of the ACF detector. Given an input image $I$, we compute several channels $C = \Omega(I)$, sum every block of pixels in $C$, and smooth the resulting lower resolution channels. Features are single pixel lookups in the aggregated channels. Boosting is used to learn decision trees over these features (pixels) to distinguish object from background. With the appropriate choice of channels and careful attention to design, ACF achieves state-of-the-art performance in pedestrian detection.

(Images from Dollár et al. 2014)
Technical approach

• Training: Train 3DVP Detectors
  • SVM-based detectors for KITTI (Malisiewicz et al. 2011)
  • Boosting detector for KITTI
    • Aggregated Channel Features (ACF) (Dollár et al. 2014)

• A trick: Incorporate the appearance of the occluder

(Figures from Xiang et al. 2015)
Technical approach

• Testing: Get 2D detection bounding boxes

(Figures from Xiang et al. 2015)
Technical approach

• Testing: Transfer the meta-data associated with the 3DVPs

(Figures from Xiang et al. 2015)
Technical approach

- Testing: Transfer the meta-data associated with the 3DVPs
  - Energy-based conditional random field model
    - $m_i = m_i^v + m_i^o + m_i^t$ (visible, occluded, and truncated)
Technical approach

- **Testing**: Transfer the meta-data associated with the 3DVPs
  - Energy-based conditional random field model
    - $m_i = m_i^v + m_i^o + m_i^t$ (visible, occluded, and truncated)

\[
E(\hat{D}) = \sum_{i \in \hat{D}} \left( \frac{w_d(s_i - b)}{\text{detection score}} - w_o \frac{|m_i^o| + |m_i^t|}{|m_i|} + w_o \frac{|m_i^t \notin I|}{|m_i|} \right) + \\
\sum_{i,j \in \hat{D}, i \neq j} \left( w_o \frac{|m_{\text{far}(i,j)} \cap m_{\text{near}(i,j)}^v|}{|m_{\text{far}(i,j)}|} - w_p \frac{\sum_{k=v,o,t} |m_i^k \cap m_j^k|}{\min(|m_i|, |m_j|)} \right)
\]

- **Implementation**: Greedy algorithm


Technical approach

• Testing: Transfer the meta-data associated with the 3DVPs
  • Non–Maximum Suppression (NMS) (Felzenszwalb et al. 2010)
    • Sort the results, and pick the one with largest score
    • Computes the overlap between two bounding boxes by $\frac{|a_i \cap a_j|}{|a_i|}$
    • Greedily suppress detections that have larger than 0.5 overlap with selected ones
    • Noted by “NMS.5” in this paper
  • Intersection over Union (IoU) with 0.6 threshold
    • NMS-based, but keep more occluded detection hypotheses
    • Noted by “INMS.6” in this paper
Experimental evaluation

• Datasets
  • KITTI:
    • 7481 images (28,612 cars)
    • Split the training set into training set and validation set
  • OutdoorScene:
    • 200 images (focus on the presence of severe occlusions)
    • Only for testing
Experimental evaluation

• Evaluation metrics (threshold based metrics)
  • Object detection: Average Precision (AP) (Everingham et al. 2011)
  • Object orientation: Average Orientation Similarity (AOS) (Geiger et al. 2012)

\[
AOS = \frac{1}{11} \sum_{r \in \{0,0.1,...,1\}} \max_{\tilde{r}: \tilde{r} \geq r} s(\tilde{r})
\]

where

\[
r = \frac{TP}{TP + FN} \quad s(r) = \frac{1}{|D(r)|} \sum_{i \in D(r)} \frac{1 + \cos \Delta_{\theta}^{(i)}}{2} \delta_i \quad \in [0,1]
\]
Experimental evaluation

- Evaluation metrics (threshold based metrics)
  - Object detection: Average Precision (AP) (Everingham et al. 2011)
  - Object orientation: Average Orientation Similarity (AOS) (Geiger et al. 2012)

\[
AOS = \frac{1}{11} \sum_{r \in \{0,0.1,...,1\}} \max_{\bar{r}: \bar{r} \geq r} s(\bar{r})
\]

where

\[
r = \frac{TP}{TP+FN}, \quad s(r) = \frac{1}{|D(r)|} \sum_{i \in D(r)} \frac{1 + \cos \Delta_{si}}{2} \delta_i \quad \in [0,1]
\]
Experimental evaluation

• Evaluation metrics (threshold based metrics)
  • Object detection: Average Precision (AP) (Everingham et al. 2011)
  • Object orientation: Average Orientation Similarity (AOS) (Geiger et al. 2012)

\[
AOS = \frac{1}{11} \sum_{r \in \{0, 0.1, ..., 1\}} \max_{\tilde{r} : \tilde{r} \geq r} s(\tilde{r})
\]

where

\[
r = \frac{TP}{TP + FN} \quad s(r) = \frac{1}{|D(r)|} \sum_{i \in D(r)} \frac{1 + \cos \Delta_{\theta}^{(i)}}{2} \delta_i \in [0, 1]
\]

• 2D segmentation: Average Segmentation Accuracy (ASA)
• 3D localization: Average Localization Precision (ALP)
Experimental evaluation

• Result: 2D clustering vs 3D clustering

<table>
<thead>
<tr>
<th>2D K-means</th>
<th>3D K-means</th>
<th>2D Affinity Propagation</th>
<th>3D Affinity Propagation</th>
</tr>
</thead>
<tbody>
<tr>
<td>K</td>
<td>Easy</td>
<td>Moderate</td>
<td>Hard</td>
</tr>
<tr>
<td>5</td>
<td>44.21</td>
<td>31.23</td>
<td>25.42</td>
</tr>
<tr>
<td>10</td>
<td>47.78</td>
<td>38.13</td>
<td>32.26</td>
</tr>
<tr>
<td>20</td>
<td>61.24</td>
<td>48.04</td>
<td>40.27</td>
</tr>
<tr>
<td>30</td>
<td>67.83</td>
<td>51.68</td>
<td>43.63</td>
</tr>
<tr>
<td>40</td>
<td>66.49</td>
<td>53.18</td>
<td>45.96</td>
</tr>
<tr>
<td>50</td>
<td>66.65</td>
<td>54.90</td>
<td>43.38</td>
</tr>
<tr>
<td>100</td>
<td>58.45</td>
<td>46.15</td>
<td>39.34</td>
</tr>
<tr>
<td>150</td>
<td>56.74</td>
<td>43.84</td>
<td>37.75</td>
</tr>
</tbody>
</table>

Table 1. AP Comparison between 2D and 3D clustering with k-means and affinity propagation on our validation split. The table shows the average precision obtained by training ACF detectors in different settings.

(Table from Xiang et al. 2015)
Experimental evaluation

• Result: Occlusion (Energy-based) vs NMS.5 vs INMS.6
  • DPM: baselines (Felzenszwalb et al. 2010)

<table>
<thead>
<tr>
<th>Methods</th>
<th>Object Detection (AP)</th>
<th>Orientation (AOS)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Easy</td>
<td>Moderate</td>
</tr>
<tr>
<td>DPM [10] NMS.5</td>
<td>54.91</td>
<td>42.49</td>
</tr>
<tr>
<td>DPM [10] INMS.6</td>
<td>44.35</td>
<td>36.49</td>
</tr>
<tr>
<td>Ours NMS.5</td>
<td>79.06</td>
<td>64.72</td>
</tr>
<tr>
<td>Ours INMS.6</td>
<td>78.28</td>
<td>65.62</td>
</tr>
<tr>
<td>Ours Occlusion</td>
<td><strong>80.48</strong></td>
<td><strong>68.05</strong></td>
</tr>
</tbody>
</table>

Table 2. AP/AOS comparison between different detection/decoding methods on the validation set. We show the results of 3D AP with 125 clusters for Ours.

(Table from Xiang et al. 2015)
Experimental evaluation

• Result: 2D segmentation
  • Lack of ground truth: projecting registered 3D CAD models

<table>
<thead>
<tr>
<th>Method</th>
<th>Easy</th>
<th>Moderate</th>
<th>Hard</th>
</tr>
</thead>
<tbody>
<tr>
<td>DPM [10]+box</td>
<td>38.09</td>
<td>29.42</td>
<td>22.65</td>
</tr>
<tr>
<td>Ours INMS.6+box</td>
<td>57.52</td>
<td>47.84</td>
<td>40.01</td>
</tr>
<tr>
<td>Ours Occlusion+box</td>
<td>59.21</td>
<td>49.74</td>
<td>41.71</td>
</tr>
<tr>
<td>Ours INMS.6+3DVP</td>
<td>63.88</td>
<td>52.57</td>
<td>43.82</td>
</tr>
<tr>
<td>Ours Occlusion+3DVP</td>
<td>65.73</td>
<td>54.60</td>
<td>45.62</td>
</tr>
</tbody>
</table>

(Table from Xiang et al. 2015)
Experimental evaluation

• Result: 2D segmentation
• Qualitative result:

(Images from Xiang et al. 2015)
Experimental evaluation

• Result: 3D localization

(Table from Xiang et al. 2015)
Experimental evaluation

• Result: 3D localization
  • Qualitative result:

Figure 8. 2D recognition and 3D localization results on the KITTI test set. Blue regions in the images are the estimated occluded areas.

(Images and videos from Xiang et al. 2015)
Experimental evaluation

• Result: KITTI test set evaluation
  • Use the whole training set to generate the 3DVPs

<table>
<thead>
<tr>
<th>Methods</th>
<th>Object Detection (AP)</th>
<th>Orientation (AOS)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Easy</td>
<td>Moderate</td>
</tr>
<tr>
<td>ACF [8]</td>
<td>55.89</td>
<td>54.74</td>
</tr>
<tr>
<td>DPM [10]</td>
<td>71.19</td>
<td>62.16</td>
</tr>
<tr>
<td>DPM-VOC+VP [29]</td>
<td>74.95</td>
<td>64.71</td>
</tr>
<tr>
<td>OC-DPM [30]</td>
<td>74.94</td>
<td>65.95</td>
</tr>
<tr>
<td>SubCat [27]</td>
<td>81.94</td>
<td>66.32</td>
</tr>
<tr>
<td>AOG [24]</td>
<td>84.36</td>
<td>71.88</td>
</tr>
<tr>
<td>SubCat [28]</td>
<td>84.14</td>
<td>75.46</td>
</tr>
<tr>
<td>Regionlets [36]</td>
<td>84.75</td>
<td>76.45</td>
</tr>
<tr>
<td><strong>Ours</strong> INMS.6</td>
<td>84.81</td>
<td>73.02</td>
</tr>
<tr>
<td><strong>Ours</strong> Occlusion</td>
<td><strong>87.46</strong></td>
<td><strong>75.77</strong></td>
</tr>
</tbody>
</table>

Table 4. AP/AOS Comparison between different methods on the KITTI test set. We show the results of 3D AP with 227 clusters for **Ours**. More comparisons are available at [16].

(Table from Xiang et al. 2015)
Experimental evaluation

• Result: OutdoorScene dataset evaluation
  • 3DVP detectors are generalizable to other scenarios

<table>
<thead>
<tr>
<th>% occlusion</th>
<th>&lt; 0.3</th>
<th>0.3 – 0.6</th>
<th>&gt; 0.6</th>
</tr>
</thead>
<tbody>
<tr>
<td># images</td>
<td>66</td>
<td>68</td>
<td>66</td>
</tr>
<tr>
<td>ALM [40]</td>
<td>72.3</td>
<td>42.9</td>
<td>35.5</td>
</tr>
<tr>
<td>DPM [10]</td>
<td>75.9</td>
<td>58.6</td>
<td>44.6</td>
</tr>
<tr>
<td>SLM [41]</td>
<td>80.2</td>
<td>63.3</td>
<td>52.9</td>
</tr>
<tr>
<td>Ours NMS.5</td>
<td>89.7</td>
<td>76.3</td>
<td>55.9</td>
</tr>
<tr>
<td>Ours Occlusion</td>
<td><strong>90.0</strong></td>
<td><strong>76.5</strong></td>
<td><strong>62.1</strong></td>
</tr>
</tbody>
</table>

Table 5. AP of the car detection on the OutdoorScene dataset [41].

(Table from Xiang et al. 2015)
Discussion

• **Strength of the approach**
  • Estimate detailed properties of objects beyond 2D bounding boxes

• **Weakness of the approach**
  • Running time: not mentioned in this paper
  • KITTI website
Discussion

<table>
<thead>
<tr>
<th>Method</th>
<th>Setting</th>
<th>Code</th>
<th>Moderate</th>
<th>Easy</th>
<th>Hard</th>
<th>Runtime</th>
<th>Environment</th>
</tr>
</thead>
<tbody>
<tr>
<td>SubCN</td>
<td></td>
<td></td>
<td>87.88 %</td>
<td>90.49 %</td>
<td>77.10 %</td>
<td>2 s</td>
<td>GPU @ 3.5 Ghz (Python + C/C++)</td>
</tr>
<tr>
<td>DML</td>
<td></td>
<td></td>
<td>87.51 %</td>
<td>90.67 %</td>
<td>76.33 %</td>
<td>x s</td>
<td>GPU @ 1.5 Ghz (Matlab + C/C++)</td>
</tr>
<tr>
<td>3DOP</td>
<td>code</td>
<td></td>
<td>86.10 %</td>
<td>91.44 %</td>
<td>76.52 %</td>
<td>3 s</td>
<td>GPU @ 2.5 Ghz (Matlab + C/C++)</td>
</tr>
<tr>
<td>Mono3D</td>
<td></td>
<td></td>
<td>85.66 %</td>
<td>88.31 %</td>
<td>75.89 %</td>
<td>x s</td>
<td>GPU @ 2.5 Ghz (Matlab + C/C++)</td>
</tr>
<tr>
<td>3DYP</td>
<td></td>
<td></td>
<td>74.59 %</td>
<td>86.92 %</td>
<td>64.11 %</td>
<td>40 s</td>
<td>8 cores @ 3.5 Ghz (Matlab + C/C++)</td>
</tr>
<tr>
<td>SubCat</td>
<td>code</td>
<td></td>
<td>74.42 %</td>
<td>83.41 %</td>
<td>58.83 %</td>
<td>0.7 s</td>
<td>6 cores @ 3.5 Ghz (Matlab + C/C++)</td>
</tr>
<tr>
<td>SubCat+HSC</td>
<td></td>
<td></td>
<td>73.95 %</td>
<td>83.07 %</td>
<td>58.29 %</td>
<td>5.5 s</td>
<td>2 cores @ 2.5 Ghz (Matlab + C++)</td>
</tr>
<tr>
<td>SS</td>
<td></td>
<td></td>
<td>73.06 %</td>
<td>83.87 %</td>
<td>58.38 %</td>
<td>0.3 s</td>
<td>4 cores @ 2.5 Ghz (Matlab + C/C++)</td>
</tr>
<tr>
<td>SubCat</td>
<td></td>
<td></td>
<td>64.94 %</td>
<td>80.92 %</td>
<td>50.03 %</td>
<td>0.3 s</td>
<td>6 cores @ 2.5 Ghz (Matlab + C/C++)</td>
</tr>
<tr>
<td>OC-DPM</td>
<td></td>
<td></td>
<td>64.42 %</td>
<td>73.50 %</td>
<td>52.40 %</td>
<td>10 s</td>
<td>8 cores @ 2.5 Ghz (Matlab)</td>
</tr>
</tbody>
</table>

(Screenshot from KITTI website: Geiger et al. 2012)
Discussion

• Strength of the approach
  • Estimate detailed properties of objects beyond 2D bounding boxes

• Weakness of the approach
  • Running time: not mentioned in this paper
  • KITTI website

• Future direction
  • Be able to adapt to different problems using different CAD models (e.g., Cyclists, Pedestrians)
3D Object Proposals for Accurate Object Class Detection

Xiaozhi Chen*,1  Kaustav Kundu *,2  Yukun Zhu2  Andrew Berneshawi2  
Huimin Ma1  Sanja Fidler2  Raquel Urtasun2  

1Department of Electronic Engineering  
Tsinghua University  

2Department of Computer Science  
University of Toronto
High-level Overview

• Propose a new object proposal approach: 3D object proposals (3DOP)
  • In the context of autonomous driving
  • Exploits stereo imagery to place 3D bounding boxes

• Complete the full pipeline combing 3DOP and CNN

(Images from Chen et al. 2015)
High-level Overview

• Propose a new object proposal approach: 3D object proposals (3DOP)
  • In the context of autonomous driving
  • Exploits stereo imagery to place 3D bounding boxes
• Complete the full pipeline combing 3DOP and CNN
• Experiments on KITTI benchmark
  • Outperforms all existing approaches on all three categories (cars, cyclists, and pedestrians)
Motivation

• Why generating the proposal before object detection?
  • Proposals: at least a few accurately cover the ground-truth objects
  • Split the system into two phases:
    • i) generate the image proposals and ii) classify each proposal
• Combine with other algorithm like R-CNN
  • Challenging conditions in autonomous driving
Motivation

• Why generating the proposal before object detection?
  • Proposals: at least a few accurately cover the ground-truth objects
  • Split the system into two phases:
    • i) generate the image proposals and ii) classify each proposal
  • Combine with other algorithm like R-CNN
    • Challenging conditions in autonomous driving

• Inspired by previous work
  • Selective Search (Van de Sande et al. 2011)
  • Contours-based method (Zitnick and Dollár 2014)
Motivation

• Challenges
  • High computational complexity of sliding windows
  • Produce perfect recall with fewer proposals
    • Trade-off between recall rate and precision rate
  • Exploit the stereo imagery to improve the performance
Technical approach

• Proposal Generation as Energy Minimization

\[ E(x, y) = \mathbf{w}_{c, pcd}^\top \phi_{pcd}(x, y) + \mathbf{w}_{c, fs}^\top \phi_{fs}(x, y) + \mathbf{w}_{c, ht}^\top \phi_{ht}(x, y) + \mathbf{w}_{c, ht-contr}^\top \phi_{ht-contr}(x, y) \]

• \( x \): point cloud
• \( y \): tuple \((x, y, z, \theta, c, t)\)
• \( \mathbf{w}_c^\top \): class-specific weights
Technical approach

• Proposal Generation as Energy Minimization

\[ E(x, y) = w_{c,pcd}^T \phi_{pcd}(x, y) + w_{c,fs}^T \phi_{fs}(x, y) + w_{c,ht}^T \phi_{ht}(x, y) + w_{c,ht-contr}^T \phi_{ht-contr}(x, y) \]

• Point cloud density

\[ \phi_{pcd}(x, y) = \frac{\sum_{p \in \Omega(y)} S(p)}{|\Omega(y)|} \]

(Image from Chen et al. 2015)
Technical approach

- Proposal Generation as Energy Minimization

\[ E(x, y) = w_{c,pcd}^\top \phi_{pcd}(x, y) + w_{c,fs}^\top \phi_{fs}(x, y) + w_{c,ht}^\top \phi_{ht}(x, y) + w_{c,ht-contr}^\top \phi_{ht-contr}(x, y) \]

- Free space

\[ \phi_{fs}(x, y) = \frac{\sum_{p \in \Omega(y)} (1 - F(p))}{|\Omega(y)|} \]

(Image from Chen et al. 2015)
Technical approach

• Proposal Generation as Energy Minimization

\[ E(x, y) = w_{c,pcd}^T \phi_{pcd}(x, y) + w_{c,fs}^T \phi_{fs}(x, y) + w_{c,ht}^T \phi_{ht}(x, y) + w_{c,ht-contr}^T \phi_{ht-contr}(x, y) \]

• Height prior

\[ \phi_{ht}(x, y) = \frac{1}{|\Omega(y)|} \sum_{p \in \Omega(y)} H_c(p) \]

with

\[ H_c(p) = \begin{cases} 
\exp \left[ -\frac{1}{2} \left( \frac{d_p - \mu_{c,ht}}{\sigma_{c,ht}} \right)^2 \right], & \text{if } S(p) = 1 \\
0, & \text{o.w.}
\]
Technical approach

• Proposal Generation as Energy Minimization

\[ E(x, y) = w_{c,pcd}^T \phi_{pcd}(x, y) + w_{c,fs}^T \phi_{fs}(x, y) + w_{c,ht}^T \phi_{ht}(x, y) + w_{c,ht-ctr}^T \phi_{ht-ctr}(x, y) \]

• Height contrast

\[ \phi_{ht-ctr}(x, y) = \frac{\phi_{ht}(x, y)}{\phi_{ht}(x, y^+) - \phi_{ht}(x, y)} \]
Technical approach

• Proposal Generation as Energy Minimization

\[ E(x, y) = w_{c, pcd}^T \phi_{pcd}(x, y) + w_{c, fs}^T \phi_{fs}(x, y) + w_{c, ht}^T \phi_{ht}(x, y) + w_{c, ht-contr}^T \phi_{ht-contr}(x, y) \]

• Inference

\[ y^* = \text{argmin}_y E(x, y) \]

• Get N diverse proposals
  • Sort the values of \( E(x, y) \) for all \( y \)
  • Greedy inference: pick top scoring proposal, perform NMS (Felzenszwalb et al. 2010), and iterate
Technical approach

• Proposal Generation as Energy Minimization

\[ E(x, y) = w_{c,pcd}^T \Phi_{pcd}(x, y) + w_{c,fs}^T \Phi_{fs}(x, y) + w_{c,ht}^T \Phi_{ht}(x, y) + w_{c,ht-contr}^T \Phi_{ht-contr}(x, y) \]

• Speed up tricks
  • Integral image (summed area table)
  • Skipping configurations which do not overlap with the point cloud
  • Place all our bounding boxes on the road plane
    • Sample additional proposal boxes at large locations: \( y = y_{road} \pm \sigma_{road} \)
Technical approach

• Proposal Generation as Energy Minimization

\[ E(x, y) = w_{c,pcd}^T \phi_{pcd}(x, y) + w_{c,fs}^T \phi_{fs}(x, y) + w_{c,ht}^T \phi_{ht}(x, y) + w_{c,ht-contr}^T \phi_{ht-contr}(x, y) \]

• Learn the weights \( w_c^T \) using structured SVM (Tsochantaridis et al. 2004)

  • Given \( N \) ground truth input-output pairs \( \{x^{(i)}, y^{(i)}\}_{i=1,...,N} \), solve the optimization problem:

\[
\begin{align*}
\min_{w \in \mathbb{R}^D} & \quad \frac{1}{2} \|w\|^2 + \frac{C}{N} \sum_{i=1}^{N} \xi_i \\
\text{s.t.} & \quad w^T (\phi(x^{(i)}, y) - \phi(x^{(i)}, y^{(i)})) \geq \Delta(y^{(i)}, y) - \xi_i, \quad \forall y \neq y^{(i)}
\end{align*}
\]
Technical approach

• Object Detection and Orientation Estimation Network
  • 3DOP is combined with Fast R-CNN (Girshick 2015)

(Figure from Girshick 2015)
Technical approach

- **Object Detection and Orientation Estimation Network**
  - 3DOP is combined with Fast R-CNN (Girshick 2015)
  - A context branch after the last convolutional layer
    - Enlarging the candidate regions by a factor of 1.5 (Zhu *et al.* 2015)

(Figures from Chen *et al.* 2015)
Technical approach

• Object Detection and Orientation Estimation Network
  • 3DOP is combined with Fast R-CNN (Girshick 2015)
  • A context branch after the last convolutional layer
    • Enlarging the candidate regions by a factor of 1.5 (Zhu et al. 2015)
  • Orientation regression loss
    • Jointly learn object location and orientation
    • Smooth $L_1$ loss: Less sensitive to outliers than L2 loss used in R-CNN (Girshick et al. 2014) and SPPnet (He et al. 2015)

\[
L_{\text{loc}}(t^u, v) = \sum_{i \in \{x,y,w,h\}} \text{smooth}_{L_1}(t_i^u - v_i),
\]

in which

\[
\text{smooth}_{L_1}(x) = \begin{cases} 
0.5x^2 & \text{if } |x| < 1 \\
|x| - 0.5 & \text{otherwise,}
\end{cases}
\]
Technical approach

• Object Detection and Orientation Estimation Network
  • 3DOP is combined with Fast R-CNN (Girshick 2015)
  • A context branch after the last convolutional layer
    • Enlarging the candidate regions by a factor of 1.5 (Zhu et al. 2015)
  • Orientation regression loss
    • Jointly learn object location and orientation
    • Smooth $L_1$ loss: Less sensitive to outliers than L2 loss used in R-CNN (Girshick et al. 2014) and SPPnet (He et al. 2015)

• Initialization of weights on CNN
  • Use OxfordNet (Simonyan and Zisserman 2014) trained on ImageNet
Technical approach

• Object Detection and Orientation Estimation Network

(Figures from Chen et al. 2015)
Experimental evaluation

• Dataset: KITTI
  • 7481 training images, which contains three classes: Car, Pedestrian, and Cyclist
  • Three regimes based on the occlusion levels: Easy, Moderate, and Hard
  • Split the training set into training set and validation set

• Evaluation metric: Oracle recall (Van de Sande et al. 2011; Hosang et al. 2015)
  • For each ground truth (GT) object we found the proposal that overlaps the most in Intersection over Union (IoU)
  • Then we say it is recalled if IoU exceeds 70% for cars and 50% for pedestrians and cyclists
Experimental evaluation

- Results: Recall as a function of the number of candidates

(Figures from Chen et al. 2015)
Experimental evaluation

• Results: Recall as a function of the number of candidates

(Figures from Chen et al. 2015)
Experimental evaluation

• Results: Recall vs IoU for 500 proposals

(Figures from Chen et al. 2015)
Experimental evaluation

• Results: Running time

<table>
<thead>
<tr>
<th>Method</th>
<th>BING</th>
<th>Selective Search</th>
<th>Edge Boxes (EB)</th>
<th>MCG</th>
<th>MCG-D</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time (seconds)</td>
<td>0.01</td>
<td>15</td>
<td>1.5</td>
<td>100</td>
<td>160</td>
<td>1.2</td>
</tr>
</tbody>
</table>

Table 3: Running time of different proposal methods.

(Table from Chen et al. 2015)
Experimental evaluation

• Results: Full object detection pipeline

<table>
<thead>
<tr>
<th></th>
<th>Cars</th>
<th></th>
<th>Pedestrians</th>
<th></th>
<th>Cyclists</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Easy</td>
<td>Moderate</td>
<td>Hard</td>
<td>Easy</td>
<td>Moderate</td>
<td>Hard</td>
</tr>
<tr>
<td>LSVM-MDPM-sv [35]</td>
<td>68.02</td>
<td>56.48</td>
<td>44.18</td>
<td>47.74</td>
<td>39.36</td>
<td>35.95</td>
</tr>
<tr>
<td>SquaresICF [36]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>57.33</td>
<td>44.42</td>
<td>40.08</td>
</tr>
<tr>
<td>DPM-C8B1 [37]</td>
<td>74.33</td>
<td>60.99</td>
<td>47.16</td>
<td>38.96</td>
<td>29.03</td>
<td>25.61</td>
</tr>
<tr>
<td>MDPM-un-BB [1]</td>
<td>71.19</td>
<td>62.16</td>
<td>48.43</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DPM-VOC+VP [27]</td>
<td>74.95</td>
<td>64.71</td>
<td>48.76</td>
<td>59.48</td>
<td>44.86</td>
<td>40.37</td>
</tr>
<tr>
<td>OC-DPM [38]</td>
<td>74.94</td>
<td>65.95</td>
<td>53.86</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>AOG [39]</td>
<td>84.36</td>
<td>71.88</td>
<td>59.27</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SubCat [28]</td>
<td>84.14</td>
<td>75.46</td>
<td>59.71</td>
<td>54.67</td>
<td>42.34</td>
<td>37.95</td>
</tr>
<tr>
<td>DA-DPM [40]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>56.36</td>
<td>45.51</td>
<td>41.08</td>
</tr>
<tr>
<td>Fusion-DPM [41]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>59.51</td>
<td>46.67</td>
<td>42.05</td>
</tr>
<tr>
<td>R-CNN [42]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>61.61</td>
<td>50.13</td>
<td>44.79</td>
</tr>
<tr>
<td>FilteredICF [43]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>61.14</td>
<td>53.98</td>
<td>49.29</td>
</tr>
<tr>
<td>pAUCEnST [44]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>65.26</td>
<td>54.49</td>
<td>48.60</td>
</tr>
<tr>
<td>MV-RGBD-RE [45]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>70.21</td>
<td>54.56</td>
<td>51.25</td>
</tr>
<tr>
<td><strong>3DVP [12]</strong></td>
<td>87.46</td>
<td>75.77</td>
<td>65.38</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Regionlets [13]</td>
<td>84.75</td>
<td>64.45</td>
<td>59.70</td>
<td>73.14</td>
<td>61.15</td>
<td>55.21</td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td><strong>93.04</strong></td>
<td><strong>88.64</strong></td>
<td><strong>79.16</strong></td>
<td><strong>81.78</strong></td>
<td><strong>67.47</strong></td>
<td><strong>64.70</strong></td>
</tr>
</tbody>
</table>

Table 1: Average Precision (AP) (in %) on the test set of the KITTI Object Detection Benchmark.
Experimental evaluation

• Results: Full object orientation estimation pipeline

Table 2: AOS scores (in %) on the test set of KITTI’s Object Detection and Orientation Estimation Benchmark.

(Table from Chen et al. 2015)
Experimental evaluation

• Results: Full object orientation estimation pipeline

<table>
<thead>
<tr>
<th>Method</th>
<th>Setting</th>
<th>Code</th>
<th>Moderate</th>
<th>Easy</th>
<th>Hard</th>
<th>Runtime</th>
<th>Environment</th>
<th>Compare</th>
</tr>
</thead>
<tbody>
<tr>
<td>SubCNN</td>
<td></td>
<td></td>
<td>87.88 %</td>
<td>90.49 %</td>
<td>77.10 %</td>
<td>2 s</td>
<td>GPU @ 3.5 Ghz (Python + C/C++)</td>
<td></td>
</tr>
<tr>
<td>Anonymous submission</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DJIM</td>
<td></td>
<td></td>
<td>87.51 %</td>
<td>90.67 %</td>
<td>76.33 %</td>
<td>× s</td>
<td>GPU @ 1.5 Ghz (Matlab + C/C++)</td>
<td></td>
</tr>
<tr>
<td>3rd: 3DOP (this paper), Dec 2015, NIPS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5th: 3DVP (previous paper), June 2015, CVPR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mono3D</td>
<td></td>
<td></td>
<td>85.66 %</td>
<td>88.31 %</td>
<td>75.89 %</td>
<td>× s</td>
<td>GPU @ 3.5 Ghz (Matlab + C/C++)</td>
<td></td>
</tr>
<tr>
<td>Anonymous submission</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


(Screenshot from KITTI website: Geiger et al. 2012)
Discussion

• Strength of the approach
  • Generating proposals
    • 3DOP achieves higher recall rate on challenging KITTI benchmark
  • Full object detection/orientation estimation pipeline
    • 3DOP + Fast R-CNN outperforms state-of-the-art methods on KITTI testing set

• Weakness of the approach
  • Rely on stereo images
  • Still not a real-time algorithm (1.2 seconds for proposals, 3 seconds for full pipeline)

• Future work
  • Implement monocular 3D Object Detection
  • Improve efficiency by reducing spurious false positives


Reference


Thank you!