Beyond Bounding Boxes: Richer Object Hypotheses in Object Class Detection

Bojan Pepik
Max-Planck Institute for Informatics

- Oxford Robotics Seminar -

Joint work with Michael Stark, Peter Gehler, Tobias Ritschel and Bernt Schiele
Motivation

- **High-level vision tasks**
  - 3D scene understanding
  - Autonomous driving

- **Evidence - object detectors**
  - 2D bounding boxes

- **What do BB tell us about 3D object geometry?**
  - *Very little!*
Goal

• **Object detectors tailored towards**
  ‣ High-level applications
  ‣ 3D reasoning

• **Richer object hypotheses**
  ‣ Viewpoint
  ‣ Parts
  ‣ Sub-ordinate labels
  ‣ 3D shape
  ‣ 3D position
  ‣ Object visibility
  ‣ Contextual information
  ‣ attributes
  ‣ …

Our work

*Car 1 occludes 50% of Car 2*
Related Work

- **3D object models**
  - CAD, point clouds, depth sensors
  - Expressive
  - Not competitive detection performance

- **Multi-view object detectors**
  - Multi-class viewpoint classification
  - Limited expressiveness
  - Competitive performance

⇒ **Our goal: Expressive object model with competitive detection performance**
Related Work

• **Fine-grained recognition**
  ‣ Detailed object representations
  ‣ Discriminative local information
  ‣ Assume localized objects

• **Occlusion handling**
  ‣ Single object models
  ‣ Contextual representation

[Hariharan et al CVPR’15]
[Mothaghi et al CVPR’15]
[Krause et al CVPR’13]
[Welinder et al TR’10]

[Xiang et al CVPR’15]
[Tang et al BMVC’12]
[Wojek et al CVPR’11]
[Vedaldi et al NIPS’09]
Overview

- **Multi-view and 3D DPMs**
  [TPAMI’15, CVPR’12, ECCV’12]
  - 3D parts, continuous pose

Experiments

- Simultaneous 2D BB & VP estimation.
- RCNN FT
- Multi-view-priors-for-object-detection
- Leveraging-multiple-view-priors-for-object-detection
- Ridge-L
- KeyReg on top of an RCNN object detector rather than using the best previously published method DPM-VOC+VP-16V.

Finally, we evaluate the performance of our full 3D object class detector that predicts the precise 3D shape of the object, as confirmed by its segmentation quality. The final dataset. At the same time, it predicts the 3D shape of the object, which achieves state-of-the-art simultaneous BB localization and viewpoint estimation performance on the challenging Pascal3D+. We use the evaluation protocol of [3] and viewpoint estimation results on Pascal3D+.

In this work we have built a 3D object class detector, which outperforms all 3D object models by considerable margins. RCNN-Ridge-L achieves state-of-the-art segmentation methods (OAS) and viewpoint estimation performance on the challenging Pascal3D+ only. RCNN-KeyReg successfully predicts the 3D object shape which achieves state-of-the-art segmentation methods (OAS). In Tab. 1, we go one step further and compare to native approaches.

KeyReg) on top of an RCNN object detector rather than using the best previously published method DPM-VOC+VP-16V.

Finally, we evaluate the performance of our full 3D object class detector that predicts the precise 3D shape of the object, as confirmed by its segmentation quality. The final dataset. At the same time, it predicts the 3D shape of the object, which achieves state-of-the-art simultaneous BB localization and viewpoint estimation performance on the challenging Pascal3D+. We use the evaluation protocol of [3] and viewpoint estimation results on Pascal3D+.
Multi-view and 3D DPMs

- Deformable part model
  [Felsenzwalb et al PAMI’10]
  - we teach it 3D geometry

\[ \beta_1 \]

\[ \beta_M \]
DPM

- **Mixture of star components**
  \[ \beta = (\beta_1, \beta_2 \ldots, \beta_M) \]

- **Inference:** best matching component

![Diagram of DPM components](image)
Structured Output Formulation

- **DPM:** learn to *classify* detection hypotheses
  - Latent SVM
  - Binary output
Structured Output Formulation

- **Rank through BB localization**
  - Pascal VOC overlap criterion
    \[
    \Delta_{VOC}(y_i, \bar{y}) = 1 - \frac{y_i^b \cap \bar{y}^b}{y_i^b \cup \bar{y}^b}
    \]
    BB overlap

- **Rank through viewpoint estimation**
  - Classification error (0/1)
  - Angular distance
    \[
    \Delta_{VP}(y_i, \bar{y}) = [y_i^v \neq \bar{y}^v]
    \]
    \[
    \Delta_{VP}(y_i, \bar{y}) = \frac{\angle(y_i^v, \bar{y}^v)}{180^\circ}
    \]
Structured Output Formulation

• *Jointly* address object localization and viewpoint estimation

\[
\Delta(y, \bar{y}) = (1 - \alpha)\Delta_{VOC}(y, \bar{y}) + \alpha\Delta_{VP}(y, \bar{y})
\]

- \(\alpha \in [0, 1]\)
- \(\alpha = 0\) → DPM-VOC
- \(\alpha = 0.5\) → DPM-VOC+VP
3D Part Parameterization

- Independent set of parts across views

\[ \beta_1 \]

\[ \beta_M \]

- Part correspondences across object views
3D Part Parameterization

• Independent set of parts across views

\[ \beta_1 \quad \cdots \quad \beta_M \]

• 3D part parameterization
  ‣ 3D position
  ‣ 3D size

DPM-3D-Constraints
3D Part Parameterization

- Independent set of parts across views
- 3D part parameterization

\[
\begin{align*}
\beta_1 \quad & \quad \ldots \quad & \quad \ldots \quad & \quad \beta_M
\end{align*}
\]

- 3D position and extent
- 3D part displacement

**3D^2PM**
3D Part Parameterization

- 3D object geometry through CAD data
  - Enables learning of 3D geometry
  - Non-photo realistic rendering
  - Combined with real world examples
  - 3D part learning
Continuous Appearance Model

- Continuous appearance model
  - Linear interpolation scheme
  - Synthesize *infinitely* many viewpoints
  - *Arbitrary fine* viewpoint estimation
  - *Scalable* runtime
3D^2PM
Experiments

• **Viewpoint estimation**
  
  3D Object classes
  
  [Images of bicycles, cars, and other objects from different viewpoints]

  viewpoint classification
  
  EPFL multi-view cars
  
  [Images of cars from different viewpoints]

  angular viewpoint estimation

• **Joint object localization and viewpoint estimation**
  
  Pascal3D+
  
  [Images of a 3D model of a car and a real scene with a car]

  KITTI
  
  [Images of a real-world street scene with multiple cars]

• **Ultra-wide baseline matching experiment**
  
  [Images of cars from different wide-baseline viewpoints]
Experiments
Structured output learning

- Pascal 3D+ dataset
  - 12 categories
  - continuous viewpoint annotations
  - joint object localization and viewpoint estimation

AVP values for different viewpoint annotations:

- VP4, VP8, VP16, VP24

Graph showing performance comparison:
- Ghodrati et al. BMVC'14
- DPM-Hinge+VP
- DPM-VOC+VP

DPM-VOC+VP consistently better than DPM-Hinge+VP
Experiments
Detection performance

- KITTI testing set
  - 3 categories
  - 3D annotations
  - Detection performance

<table>
<thead>
<tr>
<th>AP</th>
<th>DPM-VOC +VP</th>
<th>DPM-3D-Const.</th>
<th>3D^2PM</th>
<th>DPM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Car</td>
<td>48.8</td>
<td>42.2</td>
<td>45.6</td>
<td>41.0</td>
</tr>
<tr>
<td>Pedestrian</td>
<td>40.8</td>
<td>36.6</td>
<td>37.4</td>
<td>34.8</td>
</tr>
<tr>
<td>Cyclist</td>
<td>28.2</td>
<td>27.5</td>
<td>27.1</td>
<td>27.3</td>
</tr>
<tr>
<td>Avg</td>
<td>39.3</td>
<td>35.4</td>
<td>36.7</td>
<td>34.4</td>
</tr>
</tbody>
</table>

⇒ All 3 models outperform DPM
3D Part Parameterization

- Quantitative analysis: Ultra-wide baseline matching experiment [Zia et al. TPAMI’13]
  - Fundamental matrix estimation
  - 140 image pairs

✓ Average improvement of 14.4%
Overview

- **Multi-view and 3D DPMs**
  [TPAMI’15, CVPR’12, ECCV’12]
  - 3D parts, continuous pose

- **3D object class detection in the wild**
  [3DSI’15, arXiv’15]
  - 3D shape, 3D pose
Approach overview

Input Image

3D object detection

Segmentation
Accuracy = 90.8
Viewpoint estimation

- **RCNN object detection**
  - Object proposals (Selective Search)
  - AlexNet
  - SVM on fc7 features

- **Multi-view RCNN (RCNN-MV)**
  - Multi-view classifier
  - Category x viewpoints
  - fc6

- **Viewpoint regression RCNN (RCNN-Reg)**
  - Ridge and lasso regression
  - Continuous viewpoint prediction
  - pool5
Keypoint detection

- **Keypoint RCNN**
  - Keypoint annotations
  - DPM keypoint proposals
  - AlexNet FT for keypoints

- **Spatial keypoint model**
  - Relative keypoint positions
  - Max pool in keypoint neighborhood
  - Viewpoint components
3D lifting

- **3D shape representation**
  - prototypical CAD models
  - 3D keypoints

- **Camera model**
  - Pinhole camera
  - 3D pose
  - 3D translation

- **3D lifting procedure**
  - joint camera and 3D shape estimation

\[
(c^*, P^*) = \arg\min_{c, P} \sum_{i} ||k^i - \tilde{k}_c^i||.
\]
Qualitative results
Experiments
Viewpoint estimation

- **Pascal3D+**
  - angular accurate viewpoint annotations

- **Joint localization and viewpoint estimation**
  - AAVP metric

⇒ **RCNN-Reg** outperforms all multi-view methods

⇒ **RCNN-Ridge-L** state-of-the-art in joint object localization and viewpoint estimation
Experiments
3D Shape Estimation

- **Segmentation evaluation**
  - mask via CAD model projection
  - segmentation accuracy

- **Baselines**
  - RCNN-L
  - RCNN-KeyReg

<table>
<thead>
<tr>
<th>Acc</th>
<th>aero</th>
<th>bike</th>
<th>boat</th>
<th>bus</th>
<th>car</th>
<th>chair</th>
<th>table</th>
<th>mbike</th>
<th>sofa</th>
<th>train</th>
<th>tv</th>
<th>avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>GT</td>
<td>58.3</td>
<td>32.0</td>
<td>57.9</td>
<td>84.9</td>
<td>79.6</td>
<td>53.5</td>
<td>63.1</td>
<td>69.3</td>
<td>64.7</td>
<td>70.5</td>
<td>80.7</td>
<td>65.0</td>
</tr>
<tr>
<td>RCNN-KeyReg</td>
<td>27.1</td>
<td>20.2</td>
<td>19.1</td>
<td>56.2</td>
<td>47.7</td>
<td>23.0</td>
<td>18.6</td>
<td>41.3</td>
<td>46.4</td>
<td>30.9</td>
<td>70.0</td>
<td>36.4</td>
</tr>
<tr>
<td>RCNN-L</td>
<td>30.3</td>
<td>22.0</td>
<td><strong>27.9</strong></td>
<td>60.5</td>
<td>44.2</td>
<td>24.9</td>
<td>24.4</td>
<td>46.3</td>
<td>41.9</td>
<td>37.5</td>
<td>45.6</td>
<td>36.9</td>
</tr>
<tr>
<td>RCNN-Ridge-L</td>
<td><strong>35.1</strong></td>
<td><strong>22.2</strong></td>
<td>26.9</td>
<td><strong>66.4</strong></td>
<td><strong>53.9</strong></td>
<td><strong>26.8</strong></td>
<td><strong>28.6</strong></td>
<td><strong>49.0</strong></td>
<td>44.8</td>
<td><strong>42.5</strong></td>
<td>58.7</td>
<td><strong>41.4</strong></td>
</tr>
</tbody>
</table>

+4.5%
Overview

- **Multi-view and 3D DPMs**
  [TPAMI’15, CVPR’12, ECCV’12]
  - 3D parts, continuous pose

- **3D object class detection in the wild**
  [3DSI’15, arXiv’15]
  - 3D shape, 3D pose

- **Occlusion patterns**
  [CVPR’13]
  - occlusion level & type

- **Fine-grained representations**
  [ICLR’14, arXiv’14, BMVC’12]
  - sub-ordinate categories & metric size
Occlusion Patterns

Motivation

- **Occlusion**
  - treat it as a *signal*

- **Occlusion patterns**
  - [first contribution]
    - object-object occlusion interactions

- **Occlusion aware detectors**
  - [second contribution]
    - single object detectors
    - double object detectors

- **Improve performance**
  - [third contribution]
    - hard occlusion cases
Fine-grained representations

- **Multi-view subordinate model**
  - Higher precision

- **Hierarchical knowledge transfer**
  - Multi-view recognition
  - Subordinate classes

- **Prior distribution over permissible models**
  - Dense cell correlations
  - Sparse cell correlations

- **Extensive study**
  - Realistic driving setup
  - Controlled setup
Conclusion

- Multi-view and 3D DPMs
  [TPAMI’15, CVPR’12, ECCV’12]
  - 3D^2PM on par with 2D detectors

- 3D object class detection in the wild
  [3DSI’15, arXiv’15]
  - RCNN-Ridge-L state-of-the-art 3D object detection method

- Occlusion patterns
  [CVPR’13]
  - OC-DPM

- Fine-grained representations
  [ICLR’14, arXiv’14, BMVC’12]
Outlook

• Richer object hypotheses and CNNs
  ‣ Architecture design
  ‣ Structured output learning with CNNs

• 3D lifting
  ‣ Richer shape representation
  ‣ Appearance factor

• CNN analysis
  ‣ 3D appearance factors
  ‣ 3D geometry

• Beyond object detection
  ‣ 3D scene understanding
  ‣ Autonomous driving
Questions