3D Object Detection and Pose Estimation

Yu Xiang
University of Michigan
1st Workshop on Recovering 6D Object Pose
12/17/2015
2D Object Detection
2D detection is NOT enough!
Applications that need 3D Object Detection

Autonomous Driving

Robotics

Virtual Reality

Gaming

Any application that interacts with the 3D world!
Goal: Infer the 3D World

A 2D image

The 3D world

• Interaction
• Control
• Decision making
• Navigation
Etc.
Our Work: 2D Object Detection
Our Work: 2D Object Detection
Our Work: 2D Object Segmentation
Our Work: Occlusion Reasoning
Our Work: Occlusion Reasoning
Our Work: 3D Localization
Contribution: 3D Object Representations

3D Object Representation

A 2D image

The 3D world
Related Work: 2D Object Representations

- **Deformable part model**
  - Felzenszwalb et al., TPAMI’10

- 2D detection
- 3D pose
- Occlusion
- 3D location

- Viola & Jones, IJCV’01
- Fergus et al., CVPR’03
- Leibe et al., ECCVW’04
- Hoiem et al., CVPR’06
- Vedaldi et al., ICCV’09
- Maji & Malik, CVPR’09
- Felzenszwalb et al., TPAMI’10
- Malisiewicz et al., ICCV’11
- Divvala et al., ECCVW’12
- Dolla’r et al., TPAMI’14
- Etc.
Related Work: 2.5D Object Representations

- Savarese & Fei-Fei ICCV’07

- Thomas et al., CVPR’06
- Savarese & Fei-Fei ICCV’07
- Kushal et al., CVPR’07
- Su et al., ICCV’09
- Sun et al., CVPR’10
- Etc.

- 2D detection
- 3D pose
- Occlusion
- 3D location
Related Work: 3D Object Representations

- Yan et al., ICCV’07
- Hoiem et al., CVPR’07
- Liebelt et al., CVPR’08, 10
- Glasner et al. ICCV’11
- Pepik et al., CVPR’12
- Xiang & Savarese, CVPR’12
- Hejrati & Ramanan, NIPS’12
- Fidler et al., NIPS’12
- Etc
Contribution: 3D Object Representations

- 2D detection
- 3D pose
- Occlusion
- 3D location
Outline

• 3D Aspect Part Representation

• 3D Aspectlet Representation

• 3D Voxel Pattern Representation

• Conclusion and Future Work
Outline

• 3D Aspect Part Representation

• 3D Aspectlet Representation

• 3D Voxel Pattern Representation

• Conclusion and Future Work
3D Aspect Part Representation

Viewpoint Variation
3D Aspect Part Representation

Viewpoint: Azimuth 315°, Elevation 30°, Distance 2

3D Aspect Parts from 3D CAD Models

Mean Shape
3D Aspect Part Representation

Bicycle  Car  Cellphone  Iron  Mouse  Shoe

Stapler  Toaster  Bed  Chair  Sofa  Table
Aspect Layout Model

An input image

3D aspect part representation

Viewpoint: Azimuth 315°, Elevation 30°, Distance 2

Output
Aspect Layout Model

- Posterior distribution

$$P(Y, L, O, V | I) \propto \exp(E(Y, L, O, V, I))$$

$L = (l_1, ..., l_n), \ l_i = (x_i, y_i)$
Aspect Layout Model

• Energy function

\[ E(Y,L,O,V,I) = \begin{cases} 
\sum_{i} V_1(l_i,O,V,I) + \sum_{(i,j)} V_2(l_i,l_j,O,V), & \text{if } Y = +1 \\
0, & \text{if } Y = -1 
\end{cases} \]

unary potential  pairwise potential
Aspect Layout Model

• Unary potential

\[ V_1(l_i, O, V, I) = \begin{cases} 
  w_i^T \phi(l_i, O, V, I), & \text{if unoccluded} \\
  \alpha_i, & \text{if self-occluded}
\end{cases} \]
Aspect Layout Model

\[ V_1(l_i, O, V, I) = \begin{cases} 
    w_i^T \phi(l_i, O, V, I), & \text{if unoccluded} \\
    \alpha_i, & \text{if occluded}
\end{cases} \]
Aspect Layout Model

• Pairwise potential

\[ V_2(l_i, l_j, O, V) = -w_x (x_i - x_j + d_{ij,O,V} \cos(\theta_{ij,O,V}))^2 - w_y (y_i - y_j + d_{ij,O,V} \sin(\theta_{ij,O,V}))^2 \]

3D world

2D projection

2D observation
Aspect Layout Model

• Training with Structural SVM [1]

$$\min_\theta \frac{1}{2} \|\theta\|^2 + \lambda \sum_{t=1}^{N} \left[ \max_{Y,L,O,V} \left[ \theta^T \Psi_{t,Y,L,O,V} + \Delta_{t,Y,L,O,V} \right] - \theta^T \Psi_{t,Y^t,L^t,O^t,V^t} \right]$$

• Inference \((Y^*, L^*, O^*, V^*) = \arg \max_{Y,L,O,V} E(Y, L, O, V, I|\theta)\)
  
  • Loop over discretized viewpoints
  
  • Run Belief Propagation [2] under each viewpoint to predict part locations


Aspect Layout Model

• Best results upon publication in pose estimation and 3D part estimation

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cars from 3D Object dataset [Savarese &amp; Fei-Fei ICCV’07]</td>
<td>Viewpoint (cars)</td>
<td>93.4%</td>
<td>85.4</td>
<td>85.3</td>
<td>81</td>
<td>70</td>
<td>67</td>
<td>48.5</td>
</tr>
<tr>
<td>Cars from EPFL dataset [Ozuysal et al. CVPR’09]</td>
<td>Viewpoint (cars)</td>
<td>64.8%</td>
<td>58.1</td>
<td>56.6</td>
<td>41.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chairs, tables, sofas and beds from IMAGE NET [Deng et al. CVPR’09]</td>
<td>Viewpoint</td>
<td>63.4%</td>
<td>34.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Aspect Layout Model

Prediction: \( a=225, e=30, d=7 \)

Prediction: \( a=330, e=15, d=7 \)

Prediction: \( a=150, e=15, d=7 \)

Prediction: \( a=300, e=45, d=23 \)

Prediction: \( a=45, e=90, d=5 \)

Prediction: \( a=240, e=45, d=11 \)
Aspect Layout Model

Prediction: $a=30, e=15, d=2.5$

Prediction: $a=345, e=15, d=3.5$

Prediction: $a=315, e=30, d=2$

Prediction: $a=60, e=15, d=2$

Prediction: $a=0, e=15, d=1.5$

Prediction: $a=0, e=30, d=7$

Prediction: $a=60, e=15, d=2$

ImageNet dataset [Deng et al. 2010]
Wrong examples

Prediction: $a=45$, $e=15$, $d=1.5$

Prediction: $a=225$, $e=30$, $d=7$

Prediction: $a=0$, $e=30$, $d=7$

Prediction: $a=345$, $e=15$, $d=2.5$
Application I: Object Co-detection with 3D Aspect Parts

Application II: Multiview Object Tracking with 3D Aspect Parts

Azimuth=315.48
Elevation=4.56
Distance=4.98

Azimuth=1.34
Elevation=2.78
Distance=6.58

Azimuth=89.12
Elevation=3.73
Distance=2.34

Azimuth=25.17
Elevation=3.60
Distance=3.56

Application II: Multiview Object Tracking with 3D Aspect Parts

Outline

• 3D Aspect Part Representation

• 3D Aspectlet Representation

• 3D Voxel Pattern Representation

• Conclusion and Future Work
Occlusion in Object Recognition

Occlusion changes the appearances of objects.
3D Aspectlet Representation
3D Aspectlet Representation
## Object Detection Experiments

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Outdoor-scene</th>
<th>Indoor-scene</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>% occlusion</td>
<td></td>
</tr>
<tr>
<td></td>
<td>&lt; 0.3</td>
<td>0.3 – 0.6</td>
</tr>
<tr>
<td></td>
<td>&lt; 0.2</td>
<td>0.2-0.4</td>
</tr>
<tr>
<td></td>
<td># images</td>
<td></td>
</tr>
<tr>
<td>Outdoor-scene</td>
<td>66</td>
<td>68</td>
</tr>
<tr>
<td>Indoor-scene</td>
<td>77</td>
<td>111</td>
</tr>
<tr>
<td>ALM [1]</td>
<td>72.3</td>
<td>42.9</td>
</tr>
<tr>
<td>DPM [2]</td>
<td>75.9</td>
<td>58.6</td>
</tr>
<tr>
<td>Ours 3D Aspectlets</td>
<td><strong>80.2</strong></td>
<td><strong>63.3</strong></td>
</tr>
<tr>
<td></td>
<td><strong>45.9</strong></td>
<td><strong>34.5</strong></td>
</tr>
</tbody>
</table>

Object Detection Experiments

Outdoor Scenes

Indoor Scenes
Outline

• 3D Aspect Part Representation

• 3D Aspectlet Representation

• 3D Voxel Pattern Representation

• Conclusion and Future Work
What are the 3D aspect parts for aeroplane and bottle?
Data-Driven 3D Voxel Patterns

Training Pipeline Overview

1. Align 2D images with 3D CAD models

2. 3D voxel exemplars

3. 3D voxel patterns

4. Training 3D voxel pattern detectors
1. Align 2D Images with 3D CAD Models

2. Building 3D Voxel Exemplars

Depth ordering

2D mask labeling

3D CAD model

Voxelization

self-occluded

truncated

visible

occluded

3D voxel model

occluded

visible

truncated
2. Building 3D Voxel Exemplars

A 3D voxel exemplar \( E_i = (I_i, M_i, V_i) \)
3. Discovering 3D Voxel Patterns

3D Voxel Exemplars

Clustering in 3D voxel space

3D Voxel Patterns (3DVPs)
4. Training 3D Voxel Pattern detectors

- Train a ACF detector for each 3DVP.

4. Training 3D Voxel Pattern detectors

- Train a Convolutional Neural Network (CNN) for 3DVPs.

Under review
Testing Pipeline Overview

1. Apply 3DVP detectors

2. Transfer meta-data
   3. Occlusion reasoning

4. Backproject to 3D

Input 2D image

2D detection

2D segmentation
1. Apply 3DVP Detectors
1. Apply 3DVP Detectors
2. Transfer Meta-Data

3D Voxel Patterns
2. Transfer Meta-Data
3. Occlusion Reasoning

Occlusion reasoning: find a set of visibility-compatible detections

\[ E = \sum_i (\psi_{\text{detection\_score}} + \psi_{\text{truncation}}) + \sum_{ij} \psi_{\text{occlusion}} \]
3. Occlusion Reasoning
3. Occlusion Reasoning
4. 3D Localization

Backprojection
## Car Detection and Orientation Estimation on KITTI

<table>
<thead>
<tr>
<th>Method</th>
<th>Object Detection (AP)</th>
<th>Object Detection and Orientation estimation (AOS)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Easy</td>
<td>Moderate</td>
</tr>
<tr>
<td>ACF [1]</td>
<td>55.89</td>
<td>54.77</td>
</tr>
<tr>
<td>DPM [2]</td>
<td>68.02</td>
<td>56.48</td>
</tr>
<tr>
<td>DPM-VOC+VP [3]</td>
<td>74.95</td>
<td>64.71</td>
</tr>
<tr>
<td>OC-DPM [4]</td>
<td>74.94</td>
<td>65.95</td>
</tr>
<tr>
<td>SubCat [5]</td>
<td>84.14</td>
<td>75.46</td>
</tr>
<tr>
<td>Regionlets [6]</td>
<td>84.75</td>
<td><strong>76.45</strong></td>
</tr>
<tr>
<td>AOG [7]</td>
<td>84.80</td>
<td>75.94</td>
</tr>
<tr>
<td><strong>Ours 3DVP</strong></td>
<td><strong>84.81</strong></td>
<td>73.02</td>
</tr>
</tbody>
</table>

## Car Detection and Orientation Estimation on KITTI

<table>
<thead>
<tr>
<th>Method</th>
<th>Easy</th>
<th>Moderate</th>
<th>Hard</th>
<th>Easy</th>
<th>Moderate</th>
<th>Hard</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACF [1]</td>
<td>55.89</td>
<td>54.77</td>
<td>42.98</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>DPM [2]</td>
<td>68.02</td>
<td>56.48</td>
<td>44.18</td>
<td>67.27</td>
<td>55.77</td>
<td>43.59</td>
</tr>
<tr>
<td>DPM-VOC+VP [3]</td>
<td>74.95</td>
<td>64.71</td>
<td>48.76</td>
<td>72.28</td>
<td>61.84</td>
<td>46.54</td>
</tr>
<tr>
<td>OC-DPM [4]</td>
<td>74.94</td>
<td>65.95</td>
<td>53.86</td>
<td>73.50</td>
<td>64.42</td>
<td>52.40</td>
</tr>
<tr>
<td>SubCat [5]</td>
<td>84.14</td>
<td>75.46</td>
<td>59.71</td>
<td>83.41</td>
<td>74.42</td>
<td>58.83</td>
</tr>
<tr>
<td>Regionlets [6]</td>
<td>84.75</td>
<td><strong>76.45</strong></td>
<td>59.70</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>AOG [7]</td>
<td>84.80</td>
<td>75.94</td>
<td>60.70</td>
<td>33.79</td>
<td>30.77</td>
<td>24.75</td>
</tr>
<tr>
<td><strong>Ours</strong> 3DVP</td>
<td>84.81</td>
<td>73.02</td>
<td>63.22</td>
<td>84.31</td>
<td>71.99</td>
<td>62.11</td>
</tr>
<tr>
<td><strong>Ours</strong> Occlusion</td>
<td><strong>87.46</strong></td>
<td>75.77</td>
<td><strong>65.38</strong></td>
<td><strong>86.92</strong></td>
<td><strong>74.59</strong></td>
<td><strong>64.11</strong></td>
</tr>
</tbody>
</table>

### Car Detection and Orientation Estimation on KITTI

<table>
<thead>
<tr>
<th>Method</th>
<th>Easy</th>
<th>Moderate</th>
<th>Hard</th>
<th>Easy</th>
<th>Moderate</th>
<th>Hard</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACF [1]</td>
<td>55.89</td>
<td>54.77</td>
<td>42.98</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>DPM [2]</td>
<td>68.02</td>
<td>56.48</td>
<td>44.18</td>
<td>67.27</td>
<td>55.77</td>
<td>43.59</td>
</tr>
<tr>
<td>DPM-VOC+VP [3]</td>
<td>74.95</td>
<td>64.71</td>
<td>48.76</td>
<td>72.28</td>
<td>61.84</td>
<td>46.54</td>
</tr>
<tr>
<td>OC-DPM [4]</td>
<td>74.94</td>
<td>65.95</td>
<td>53.86</td>
<td>73.50</td>
<td>64.42</td>
<td>52.40</td>
</tr>
<tr>
<td>SubCat [5]</td>
<td>84.14</td>
<td>75.46</td>
<td>59.71</td>
<td>83.41</td>
<td>74.42</td>
<td>58.83</td>
</tr>
<tr>
<td>Regionlets [6]</td>
<td>84.75</td>
<td>76.45</td>
<td>59.70</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>AOG [7]</td>
<td>84.80</td>
<td>75.94</td>
<td>60.70</td>
<td>33.79</td>
<td>30.77</td>
<td>24.75</td>
</tr>
<tr>
<td><strong>Ours</strong> 3DVP</td>
<td>84.81</td>
<td>73.02</td>
<td>63.22</td>
<td>84.31</td>
<td>71.99</td>
<td>62.11</td>
</tr>
<tr>
<td><strong>Ours</strong> Occlusion</td>
<td>87.46</td>
<td>75.77</td>
<td>65.38</td>
<td>86.92</td>
<td>74.59</td>
<td>64.11</td>
</tr>
<tr>
<td><strong>Ours</strong> CNN</td>
<td><strong>90.74</strong></td>
<td><strong>88.55</strong></td>
<td><strong>77.95</strong></td>
<td><strong>90.49</strong></td>
<td><strong>87.88</strong></td>
<td><strong>77.10</strong></td>
</tr>
</tbody>
</table>

3D Voxel Patterns from PASCAL3D+ [1]

12 Rigid Categories

Detection and Pose Estimation on PASCAL3D+

<table>
<thead>
<tr>
<th>Method</th>
<th>Detection (AP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DPM [1]</td>
<td>29.6</td>
</tr>
<tr>
<td>R-CNN [2]</td>
<td>56.9</td>
</tr>
<tr>
<td><strong>Ours</strong> CNN</td>
<td><strong>60.7</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>4 Views (AVP)</th>
<th>8 Views (AVP)</th>
<th>16 Views (AVP)</th>
<th>24 Views (AVP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VDPM [3]</td>
<td>19.5</td>
<td>18.7</td>
<td>15.6</td>
<td>12.1</td>
</tr>
<tr>
<td>DPM-VOC+VP [4]</td>
<td>24.5</td>
<td>22.2</td>
<td>17.9</td>
<td>14.4</td>
</tr>
<tr>
<td><strong>Ours</strong> CNN</td>
<td><strong>47.5</strong></td>
<td><strong>31.9</strong></td>
<td><strong>24.5</strong></td>
<td><strong>19.3</strong></td>
</tr>
</tbody>
</table>

Conclusion

• 3D aspect part representation

• 3D aspectlet representation

• 3D voxel pattern representation
Open Questions

• How to scale up and benchmark 3D object recognition and scene understanding?

• How to combine deep learning with 3D representations for recognition?

• How to utilize videos and unlabeled data for 3D recognition?

• How to interact with the 3D world (affordance, action, decision)?
Ongoing Work

• A large scale dataset for 3D object recognition with 100 categories

ashtray  coffee_maker  fork  microphone  rifle  stove  bicycle
backpack  comb  guitar  microwave  road_pole  suitcase  boat
basket  computer  hair_dryer  mouse  satellite_disk  teapot  bottle
bed  cup  hammer  paintbrush  scissors  telephone  bus
bench  desk_lamp  headphone  pan  screwdriver  toaster  car
blackboard  dishwasher  helmet  pen  shoe  toilet  chair
bookshelf  door  iron  pencil  shovel  toothbrush  diningtable
bucket  eraser  jar  piano  sign  trash_bin  motorbike
cabinet  eyeglasses  kettle  pillow  skate  trophy  sofa
calculator  fan  key  plate  skateboard  tube  train
camera  faucet  keyboard  pot  slipper  vending_machine  tvmonitor
can  filing_cabinet  knife  printer  speaker  washing_machine
cap  fire_extinguisher  laptop  racket  spoon  watch

cellphone  fish_tank  lighter  refrigerator  stapler  wheelchair

clock  flashlight  mailbox  remote_control  stove  aeroplane
Ongoing Work

Images from ImageNet [1]

3D CAD models from ShapeNet [2]

Conclusion

• 3D aspect part representation

• 3D aspectlet representation

• 3D voxel pattern representation

Thank you!