Estimating the Aspect Layout of Object Categories

Yu Xiang and Silvio Savarese
University of Michigan at Ann Arbor
{yuxiang, silvio}@eecs.umich.edu
Traditional object recognition

- Uses 2D bounding boxes

**Face**
- From Viola & Jones, 01

**Human**
- From Barinova et al., 12

**Rigid object**
- From Felzenszwalb et al., 10

**Body part**
- From Ramanan & Sminchisescu, 06
Beyond 2D bounding boxes

• Model the 3D properties of objects
  • 3D pose
  • 3D part location

• More suitable for robotics, autonomous navigation and manipulation

From Saxena et al., 08
Our goals

Viewpoint: Azimuth 315°, Elevation 30°, Distance 2
Related work: joint object detection and pose estimation

- Savarese et al. 07, 08
- Ozuysal et al. 08
- Liebelt et al. 08, 10
- Xiao et al. 08
- Thomas et al. 08
- Sun et al. 09
- Su et al. 09
- Arie-Nachimson & Barsi 09
- Stark et al. 10
- Gu & Ren. 10
- Glasner et al. 11
- Payet & Todorovic 11
- Zia et al., 3DRR’11
- Pepik et al., CVPR’12
- Schels et al., CVPR’12
- Xiang and Savarese, CVPR’12

From Savarese & Fei-Fei ICCV’07

From Glasner et al. ICCV’11
Related work: 2D part-based model

Constellation Model
From Fergus et al. CVPR’03

Implicit Shape Model
From Leibe et al. ECCV’04 workshop

Deformable Part Model (DPM)

From Felzenszwalb et al. CVPR’08
Related work: 3D part-based model

From Hoiem et. al., CVPR’07

From Kushal et. al., CVPR’07

From Chiu et. al., CVPR’07

Key-view 1

Key-view 2

Key-view 3

From Sun et. al. ICCV’09
Our contributions

• Propose a 3D part based representation for object categories
• Introduce the concept of aspect parts
• Jointly solve object detection, pose estimation and aspect part localization
• Significantly improve pose estimation accuracy, evaluate rigid part localization

Aspect Part

- Parts are arbitrarily defined in previous work

- Introduce parts with geometrical and topological properties, called *aspect parts*

From Fergus et al. CVPR’03

From Felzenszwalb et al., 2010.
Aspect Part

• Our definition: a portion of the object whose 3D surface is approximately either entirely visible from the observer or entirely non-visible (i.e., self-occluded)
Aspect Part

- Our definition: a portion of the object whose 3D surface is approximately either entirely visible from the observer or entirely non-visible (i.e., self-occluded)
Aspect Part

- Our definition: a portion of the object whose 3D surface is approximately either entirely visible from the observer or entirely non-visible (i.e., self-occluded)
Aspect Part

- Our definition: a portion of the object whose 3D surface is approximately either entirely visible from the observer or entirely non-visible (i.e., self-occluded)
Aspect Part

• Our definition: a portion of the object whose 3D surface is approximately either entirely visible from the observer or entirely non-visible (i.e., self-occluded)
Aspect Part

- Our definition: a portion of the object whose 3D surface is approximately either entirely visible from the observer or entirely non-visible (i.e., self-occluded)
Aspect Part

- Our definition: a portion of the object whose 3D surface is approximately either entirely visible from the observer or entirely non-visible (i.e., self-occluded)
Aspect Part

• Examples

Bed

Car

Sofa
Aspect Part

• Related to aspect graph [1]
• Related to discriminative aspect, Farhadi et al, 07

Aspect Part

• Related to object affordance or functional part

Seat
Aspect Part

• Related to geometrical attributes of object

Horizontal surface
Vertical surface
Aspect Part

• Related to scene layout estimation

From Hedau, Hoiem & Forsyth, ECCV’10
Aspect Part

- Enables the modeling of object-human interactions

From Gupta et al., CVPR’11
Outline

• Aspect layout model
• Maximal margin parameter estimation
• Model inference
• Experiments
• Conclusion
Input & output

• Input
  – 2D image $I$

• Output
  – Object label $Y \in \{+1, -1\}$
  – Part configuration in 2D $C = (c_1, \ldots, c_n)$ $c_i = (x_i, y_i, s_i)$
Aspect Layout Model

- 3D Object $O = (o_1, \ldots, o_n)$
Aspect Layout Model

- Viewpoint representation \( V=(a,e,d) \)
- 2D part shape from 3D

Azimuth, elevation and distance
Aspect Layout Model

- Model the posterior distribution

\[ P(Y, C \mid I) \quad C = (c_1, \ldots, c_n), c_i = (x_i, y_i, s_i) \]

\[ = P(Y, L, O, V \mid I) \]

\[ L = (l_1, \ldots, l_n), \quad l_i = (x_i, y_i) \]
Aspect Layout Model

• Conditional Random Field (CRF) [1]

\[ P(Y, L, O, V \mid I) \propto \exp(E(Y, L, O, V, I)) \]

• Graph structure of the CRF

Aspect Layout Model

root nodes

part nodes
Aspect Layout Model

root nodes

part nodes
Aspect Layout Model

r1  r2  r3  r4  

p1  p2  p3

root nodes
part nodes
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

• Viewpoint invariant unary potential
  • Models part appearances

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

- Rectified HOG features

ALM only needs one template for each part across all the viewpoints.
Aspect Layout Model

• Pairwise potential
  • Constrains 2D relative locations of parts
Aspect Layout Model

• Pairwise potential

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

• Energy function

\[ E(Y, L, O, V, I \mid \theta) = \theta^T \Psi(Y, L, O, V, I) \]

– Parameters

\[ \theta = (w_{i,\forall i}, \alpha_{i,\forall i}, w_x, w_y) \]

– Linear energy function
Aspect Layout Model

• Maximal margin parameter estimation
  – Energy based learning [1]: find an energy function which outputs the maximal energy value for the correct label configuration of an object
  – Training set
    \[ T = \{(I^t, Y^t, L^t, O^t, V^t), t = 1, \ldots, N\} \]
  – Structural SVM optimization [2]

Aspect Layout Model

- Model inference

\[(Y^*, L^*, O^*, V^*) = \arg \max_{Y, L, O, V} E(Y, L, O, V, I | \theta)\]

- Run Belief Propagation (BP) [1] for each combination of \(O\) and \(V\) to obtain \(E(Y = +1, L^*, O^*, V^*)\)

- Recall the graph structure

- \(Y^* = +1\) if \(E(Y = +1, L^*, O^*, V^*) > \gamma\) (detection threshold)

---

Experiments

• Datasets
  – 3DObject dataset [1]: 10 categories, 10 instances each category
  – VOC 2006 Car dataset [2]: 921 car images
  – EPFL Car dataset [3]: 2299 images, 20 instances
  – Our new ImageNet dataset [4]: Bed (400), Chair (770), sofa (800), table (670)

Experiments

• Datasets
  – 3DObject dataset [1]: 10 categories, 10 instances each category
  – VOC 2006 Car dataset [2]: 921 car images
  – EPFL Car dataset [3]: 2299 images, 20 instances
  – Our new ImageNet dataset [4]: Bed (400), Chair (770), sofa (800), table (670)

Experiments

• Evaluation measures
  – Detection: Average Precision (AP)
  – Viewpoint: average viewpoint accuracy (the average of the elements on the main diagonal of the viewpoint confusion matrix)
  – Part localization: Percentage of Correct Parts (PCP)-recall curve
Experiments

• 3D models

Bicycle  Car  Cellphone  Iron

Mouse  Shoe  Stapler  Toaster

Bed  Chair  Sofa  Table
Experiments

• Average results for eight categories on the 3DObject dataset (8 views)

<table>
<thead>
<tr>
<th>Method</th>
<th>ALM</th>
<th>[1]</th>
<th>[2]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Viewpoint</td>
<td>80.7</td>
<td>74.2</td>
<td>57.2</td>
</tr>
<tr>
<td>Detection</td>
<td>81.8</td>
<td>n/a</td>
<td>n/a</td>
</tr>
</tbody>
</table>

Experiments

• Results on the Bicycle Category in the 3DObject dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>ALM</th>
<th>[1]</th>
<th>[2]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Viewpoint</td>
<td>91.4</td>
<td>80.8</td>
<td>75.0</td>
</tr>
<tr>
<td>Detection</td>
<td>93.0</td>
<td>n/a</td>
<td>69.8</td>
</tr>
</tbody>
</table>

Experiments

• Results on the Car Category in the 3DObject dataset

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Viewpoint</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>Detection</td>
<td>98.4</td>
<td>n/a</td>
<td>99.2</td>
<td>89.9</td>
<td>76.7</td>
<td>55.3</td>
<td>n/a</td>
</tr>
</tbody>
</table>

Experiments

• Detailed average viewpoint accuracy on the 3DObject dataset

<table>
<thead>
<tr>
<th>Category</th>
<th>Bicycle</th>
<th>Car</th>
<th>Cellphone</th>
<th>Iron</th>
<th>Mouse</th>
<th>Shoe</th>
<th>Stapler</th>
<th>Toaster</th>
</tr>
</thead>
<tbody>
<tr>
<td>DPM [1]</td>
<td>88.4</td>
<td>85.0</td>
<td>62.1</td>
<td>82.7</td>
<td>40.0</td>
<td>71.7</td>
<td>58.5</td>
<td>55.0</td>
</tr>
<tr>
<td>ALM Root</td>
<td>92.5</td>
<td>89.2</td>
<td>83.4</td>
<td>86.0</td>
<td>58.7</td>
<td>82.7</td>
<td>69.2</td>
<td>59.6</td>
</tr>
<tr>
<td>ALM Full</td>
<td>91.4</td>
<td>93.4</td>
<td>85.0</td>
<td>84.6</td>
<td>66.5</td>
<td>87.0</td>
<td>72.8</td>
<td>65.2</td>
</tr>
</tbody>
</table>

Experiments

- Effect of training set sizes for viewpoint

Experiments

• Part localization on the 3DObject dataset
Experiments

Prediction: $a=45$, $e=15$, $d=5$

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

Prediction: $a=330$, $e=15$, $d=7$

Prediction: $a=150$, $e=15$, $d=7$

Prediction: $a=60$, $e=45$, $d=7$

Prediction: $a=300$, $e=45$, $d=23$

Prediction: $a=45$, $e=90$, $d=5$

Prediction: $a=240$, $e=45$, $d=11$

Prediction: $a=300$, $e=90$, $d=15$

Prediction: $a=135$, $e=0$, $d=11$

Prediction: $a=0$, $e=60$, $d=7$

Prediction: $a=225$, $e=60$, $d=7$

Prediction: $a=300$, $e=30$, $d=15$

Prediction: $a=210$, $e=30$, $d=9$
Experiments
Experiments

• Average results on the ImageNet dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>ALM Full</th>
<th>ALM Root</th>
<th>DPM [1]</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 views</td>
<td>86.5</td>
<td>79.0</td>
<td>84.6</td>
</tr>
<tr>
<td>7 views</td>
<td>63.4</td>
<td>34.0</td>
<td>49.5</td>
</tr>
</tbody>
</table>

Experiments

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

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

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

Prediction: a=330, e=30, d=9

a=30, e=30, d=9

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

a=60, 30, d=2.5

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

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

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

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

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

Prediction: $a=330$, $e=30$, $d=9$

$a=30$, $e=30$, $d=9$

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

$a=60$, $e=30$, $d=2.5$

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

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

Prediction: $a=60$, $e=30$, $d=2.5$
Conclusion

• A new Aspect Layout Model (ALM) for object detection, pose estimation and aspect part localization.

• ALM is capable of handling large number of views, locating aspect parts and reasoning self-occlusion.

• ALM can be useful for estimating functional parts or object affordances.

• Our code and datasets are available online.
Acknowledgments

Thank you!