PanoContext
A Whole-room 3D Context Model for Panoramic Scene Understanding
by Yinda Zhang, Shuran Song, Ping Tan, Jianxiong Xiao

Presented by:
William Xie
Existing Context models

Torralba, Sinha (2001)  
Carbonetto, de Freitas & Barnard (2004)  
Torralba, Murphy, Freeman (2004)  
Rabinovich et al (2007)

Sudderth, Torralba, Wilsky, Freeman (2005)  
Heitz and Koller (2008)  
Kumar, Hebert (2005)

Desai, Ramanan, and Fowlkes (2009)  
Hoiem, Efros, Hebert (2005)

<table>
<thead>
<tr>
<th></th>
<th>aero</th>
<th>bike</th>
<th>bird</th>
<th>boat</th>
<th>bottle</th>
<th>bus</th>
<th>car</th>
<th>cat</th>
<th>chair</th>
<th>cow</th>
<th>table</th>
<th>dog</th>
<th>horse</th>
<th>mbik</th>
</tr>
</thead>
<tbody>
<tr>
<td>BB</td>
<td>.339</td>
<td>.381</td>
<td>.067</td>
<td>.099</td>
<td>.278</td>
<td>.229</td>
<td>.331</td>
<td>.146</td>
<td>.153</td>
<td>.119</td>
<td>.124</td>
<td>.066</td>
<td>.322</td>
<td>.366</td>
</tr>
<tr>
<td>context</td>
<td>.351</td>
<td>.402</td>
<td>.117</td>
<td>.114</td>
<td>.284</td>
<td>.251</td>
<td>.334</td>
<td>.188</td>
<td>.166</td>
<td>.114</td>
<td>.087</td>
<td>.078</td>
<td>.347</td>
<td>.395</td>
</tr>
</tbody>
</table>

DPM on PASCAL VOC [Felzenszwalb et al.]

Improvement on PASCAL <1.5%

Slide credit: Zhang et al.
What is this object?

Slide credit: Zhang et al.
What is this object?
What is this object?
What is this object?

Slide credit: Zhang et al.
What is this object?
What is this object?

Slide credit: Zhang et al.
Why didn’t context help?
Why didn’t context help?

Perhaps we are not using the right data
PASCAL VOC

- On average: 1.5 object classes and 2.7 object instances per image
- Average camera field of view: 40° - 60° horizontal
Human Vision

- 180° horizontal field of view
- Ability to see depth
- Ability to change viewpoint
Remedy
PanoContext
PanoContext

Input: Panorama

Output: 2D projected result

Output: 3D model

Slide credit: Zhang et al.
Input: Panorama

Output: 2D projected result

Output: 3D model

Output: 3D room exploration

Slide credit: Zhang et al.
Pipeline
Pipeline

- Vanishing point estimation for panoramas
- Room layout hypothesis generation
- 3D object hypotheses generation
- Whole-room scene hypotheses generation
- Data-driven holistic ranking
Pipeline

- Vanishing point estimation for panoramas
- Room layout hypothesis generation
- 3D object hypotheses generation
- Whole-room scene hypotheses generation
- Data-driven holistic ranking
Pipeline

• Vanishing point estimation for panoramas
• Room layout hypothesis generation
• 3D object hypotheses generation
• Whole-room scene hypotheses generation
• Data-driven holistic ranking
Pipeline

- Vanishing point estimation for panoramas
- Room layout hypothesis generation
- 3D object hypotheses generation
- Whole-room scene hypotheses generation
- Data-driven holistic ranking
Generate a pool of hypotheses

Input → Room → Object → Whole Room

Slide credit: Zhang et al.
Generate a pool of hypotheses

Input

Room

Object

Whole Room

Slide credit: Zhang et al.
Room layout hypothesis
Room layout hypothesis

Line segments detection Algorithm

Slide credit: Zhang et al.
Room layout hypothesis

Hough transform for vanishing point

Slide credit: Zhang et al.
Room layout hypothesis

Hough transform for vanishing point

Classify a vanishing direction for each line

Slide credit: Zhang et al.
VANISHING POINT

THE POINT ON THE HORIZON AT WHICH RECEEDING LINES OF PERSPECTIVE CONVERGE

Room layout hypothesis

Sample 5 line segments to generate a room layout
Room layout hypothesis

Sample 5 line segments to generate a room layout
Sample 5 line segments to generate a room layout
Room layout hypothesis

Sample 5 line segments to generate a room layout

Slide credit: Zhang et al.
Room layout hypothesis

Sample 5 line segments to generate a room layout

Slide credit: Zhang et al.
Room layout hypothesis

Sample 5 line segments to generate a room layout

Slide credit: Zhang et al.
Room layout hypothesis

Pixel-wise surface direction estimation

Slide credit: Zhang et al.
Room layout hypothesis
Room layout hypothesis

Slide credit: Zhang et al.
Room layout hypothesis

Slide credit: Zhang et al.
Room layout hypothesis

Line segments

Surface normal estimation

Slide credit: Zhang et al.
Room layout hypothesis

Consistency Score: 0.770

Slide credit: Zhang et al.
Room layout hypothesis

Consistency Score: 0.770

Slide credit: Zhang et al.
Room layout hypothesis

Consistency Score: 0.770

0.711

Line segments

Surface normal estimation

Slide credit: Zhang et al.
Room layout hypothesis

Line segments

Surface normal estimation

Consistency Score: 0.770  0.711

Slide credit: Zhang et al.
Room layout hypothesis

Consistency Score: 0.770

Slide credit: Zhang et al.
Room layout hypothesis

Consistency Score: 0.770

Slide credit: Zhang et al.
Room layout hypothesis

Consistency Score: 0.770 0.711

Slide credit: Zhang et al.
Room layout hypothesis

Consistency Score: 0.770, 0.711

Top 50 only

Slide credit: Zhang et al.
Generate a pool of hypotheses

Input

Room

Object

Whole Room

Slide credit: Zhang et al.
Cuboid detection
Cuboid detection

Fitted cuboids

Slide credit: Zhang et al.
Cuboid detection

Selective search

6 rays and 3 vanishing points

Largest IoU with the segment

Slide credit: Zhang et al.
Semantic classification

Features
- Size
- Aspect ratio & Area
- Distance to walls

Random forest

Object categories
- bed
- desk
- sofa
- chair

Slide credit: Zhang et al.
Semantic classification

Features

- Size
- Aspect ratio & Area
- Distance to walls

Random forest

Object categories

- bed
- desk
- sofa
- chair

70% Accuracy

Slide credit: Zhang et al.
Semantic classification

bed
Semantic classification

nightstand

Slide credit: Zhang et al.
Semantic classification

painting
Pairwise constraint

Object centroids

Face centroid with the closest wall

Slide credit: Zhang et al.
Generate a pool of hypotheses

Input → Room → Object → Whole Room

Slide credit: Zhang et al.
Data-driven sampling

Randomly sample a room layout

With $P(\text{layout}) \propto$ normal consistency score
Data-driven sampling

Randomly sample a room layout

With \( P(\text{layout}) \propto \) normal consistency score

Slide credit: Zhang et al.
Data-driven sampling
Data-driven sampling

Decide number of object based on prior distribution:

<table>
<thead>
<tr>
<th>Item</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>painting</td>
<td>2</td>
</tr>
<tr>
<td>bed</td>
<td>1</td>
</tr>
<tr>
<td>desk</td>
<td>1</td>
</tr>
<tr>
<td>nightst</td>
<td>1</td>
</tr>
<tr>
<td>mirror</td>
<td>1</td>
</tr>
<tr>
<td>sofa</td>
<td>1</td>
</tr>
<tr>
<td>tv</td>
<td>1</td>
</tr>
<tr>
<td>window</td>
<td>1</td>
</tr>
</tbody>
</table>
Data-driven sampling

Decide number of object sampling sequence based on prior distribution: based on bottom up scores:

<table>
<thead>
<tr>
<th>object</th>
<th>score</th>
</tr>
</thead>
<tbody>
<tr>
<td>paintin</td>
<td>2</td>
</tr>
<tr>
<td>bed</td>
<td>1</td>
</tr>
<tr>
<td>desk</td>
<td>1</td>
</tr>
<tr>
<td>nightstand</td>
<td>1</td>
</tr>
<tr>
<td>mirror</td>
<td>1</td>
</tr>
<tr>
<td>sofa</td>
<td>1</td>
</tr>
<tr>
<td>tv</td>
<td>1</td>
</tr>
<tr>
<td>window</td>
<td>1</td>
</tr>
</tbody>
</table>

Slide credit: Zhang et al.
Data-driven sampling
Sample a **bed** in empty room first...

Bottom-up score as bed

Slide credit: Zhang et al.
Data-driven sampling
Sample a **bed** in empty room first...

Randomly select one according to **bottom up** priority
rectangle detection score, semantic classifier score

Slide credit: Zhang et al.
Data-driven sampling

Then, sample a nightstand given a bed

Randomly select one according to the bottom up + pair-wise priority

mean distance to the K nearest neighbors

Slide credit: Zhang et al.
Pairwise constraint

Object centroids

Face centroid with the closest wall

Slide credit: Zhang et al.
Data-driven sampling
Keep on sampling until finishing the list…

List: bed, nightstand, painting, desk, window, painting, TV, sofa, mirror

Slide credit: Zhang et al.
Data-driven sampling

Keep on sampling until finishing the list…

List: bed, nightstand, painting, desk, window, painting, TV, sofa, mirror

Slide credit: Zhang et al.
Data-driven sampling

Keep on sampling until finishing the list...

List: bed, nightstand, painting, desk, window, painting, TV, sofa, mirror

Slide credit: Zhang et al.
Data-driven sampling

Keep on sampling until finishing the list…

List: bed, nightstand, painting, desk, window, painting, TV, sofa, mirror

Slide credit: Zhang et al.
Data-driven sampling

Keep on sampling until finishing the list...

List: bed, nightstand, painting, desk, window, painting, TV, sofa, mirror
Data-driven sampling

Keep on sampling until finishing the list...

List: bed, nightstand, painting, desk, window, painting, TV, sofa, mirror

Slide credit: Zhang et al.
List: bed, nightstand, painting, desk, window, painting, TV, sofa, mirror
Data-driven sampling

Keep on sampling until finishing the list…

Whole-room sampling is finished.

Slide credit: Zhang et al.
Holistic ranking

Learn a linear SVM for scoring and take the best

\[ f(\cdot) \]

Slide credit: Zhang et al.
Holistic ranking

Learn a linear SVM for scoring and take the best

$w^T f (\cdot)$, $N$  
$w^T f (\cdot)$, $MN$  
$w^T f (\cdot)$, $N$  
$w^T f (\cdot)$, $MN$

Slide credit: Zhang et al.
Holistic ranking

Learn a linear SVM for scoring and take the best

$$w^T f$$

Slide credit: Zhang et al.
Holistic ranking

\[ l = \left[ \Delta(y, y^*) < \epsilon \right]. \]

Binary label  Matching cost

Matching Cost

# of hypotheses

⊕ positives

⊖ negatives
Holistic feature

\[ f(x, y) = \text{bottom-up feature} + \text{top-down feature} \]
Holistic feature

\[ f(\text{bottom-up feature} + \text{top-down feature}) \]
Holistic feature

$$f \left( \text{Dataset}, \ldots \right) = \text{bottom-up feature} + \text{top-down feature}$$

Slide credit: Zhang et al.
Holistic feature

\[ f(\text{Dataset}, \text{Hypothesis}) = \text{bottom-up feature} + \text{top-down feature} \]

Dataset →

Ground Truth 1
Ground Truth 2
Ground Truth N

Hypothesis

Centroid distance, IoU, semantic type consistency

Slide credit: Zhang et al.
Holistic feature

\( f(\text{Hypothesis}) = \text{bottom-up feature} + \text{top-down feature} \)

Transformed ground truth

Dataset

A ground truth room

Slide credit: Zhang et al.
Holistic feature

\[ f(\text{Hypothesis}, \text{Dataset}) = \text{bottom-up feature} + \text{top-down feature} \]

Transformed ground truth

Pick 10 with the lowest cost

Dataset

A ground truth room

Slide credit: Zhang et al.
Final outputs
Final outputs

Slide credit: Zhang et al.
How does 3D context help?

- Helps to decide sizes of objects
- Helps to decide number of objects
- Helps to constrain relative position

DPM: Wrong relative position

Our detection

Slide credit: Zhang et al.
Context v.s. Appearance

- Context is as powerful as local appearance for detection
- Context is complementary with local appearance

Slide credit: Zhang et al.
Context v.s. Appearance

- Context is as powerful as local appearance for detection
- Context is complementary with local appearance

Slide credit: Zhang et al.
Is larger FOV helpful for room layout estimation?
Is larger FOV better for context?
My Take

- Elements of the ensemble could be valuable
- Too data driven, hard to generalize
- Future: relax the cuboid constraints, try other ways to integrate visual recognition in the pipeline
Discussion

• How can the model be generalized to other scene categories (e.g. outdoor)?

• Performance on deformable or non-axis aligned objects?

• Chairs and other non-standard layout objects?

• Indoor understanding and VQA?
Is context important in sampling and ranking?
Table 2: Object detection performance  
(a) bedroom

<table>
<thead>
<tr>
<th>object type</th>
<th>bed</th>
<th>desk</th>
<th>window</th>
<th>mirror</th>
<th>door</th>
<th>nightstand</th>
<th>wardrobe</th>
<th>cabinet</th>
<th>painting</th>
<th>tv</th>
<th>chair</th>
<th>sofa</th>
</tr>
</thead>
<tbody>
<tr>
<td>global precision (%)</td>
<td>62.16</td>
<td>40.28</td>
<td>24.00</td>
<td>28.89</td>
<td>30.65</td>
<td>27.50</td>
<td>13.89</td>
<td>0.00</td>
<td>54.79</td>
<td>25.00</td>
<td>6.15</td>
<td>0</td>
</tr>
<tr>
<td>global recall (%)</td>
<td>69.70</td>
<td>36.25</td>
<td>22.64</td>
<td>31.71</td>
<td>25.68</td>
<td>33.33</td>
<td>17.86</td>
<td>0.00</td>
<td>34.48</td>
<td>27.59</td>
<td>5.80</td>
<td>0</td>
</tr>
<tr>
<td>local precision (%)</td>
<td>63.15</td>
<td>47.89</td>
<td>22.45</td>
<td>34.78</td>
<td>29.23</td>
<td>36.36</td>
<td>16.22</td>
<td>12.50</td>
<td>57.14</td>
<td>27.03</td>
<td>11.59</td>
<td>20.00</td>
</tr>
<tr>
<td>local recall (%)</td>
<td>71.21</td>
<td>42.50</td>
<td>20.75</td>
<td>39.02</td>
<td>25.68</td>
<td>48.48</td>
<td>21.43</td>
<td>5.88</td>
<td>37.93</td>
<td>34.48</td>
<td>11.59</td>
<td>3.23</td>
</tr>
</tbody>
</table>

(b) living room

<table>
<thead>
<tr>
<th>object type</th>
<th>painting</th>
<th>door</th>
<th>cabinet</th>
<th>dining table</th>
<th>window</th>
<th>heater</th>
<th>chair</th>
<th>sofa</th>
<th>coffee table</th>
<th>end table</th>
<th>tv</th>
<th>stand</th>
</tr>
</thead>
<tbody>
<tr>
<td>global precision (%)</td>
<td>43.75</td>
<td>30.25</td>
<td>15.00</td>
<td>39.29</td>
<td>16.00</td>
<td>0.00</td>
<td>22.39</td>
<td>44.09</td>
<td>37.84</td>
<td>0.00</td>
<td>6.25</td>
<td></td>
</tr>
<tr>
<td>global recall (%)</td>
<td>44.21</td>
<td>27.69</td>
<td>9.38</td>
<td>30.56</td>
<td>8.00</td>
<td>0.00</td>
<td>11.90</td>
<td>39.05</td>
<td>33.33</td>
<td>0.00</td>
<td>4.35</td>
<td></td>
</tr>
<tr>
<td>local precision (%)</td>
<td>59.49</td>
<td>45.36</td>
<td>22.73</td>
<td>38.71</td>
<td>30.77</td>
<td>20.00</td>
<td>21.05</td>
<td>59.49</td>
<td>39.39</td>
<td>20.00</td>
<td>22.22</td>
<td></td>
</tr>
<tr>
<td>local recall (%)</td>
<td>49.47</td>
<td>33.85</td>
<td>15.63</td>
<td>33.33</td>
<td>16.00</td>
<td>16.67</td>
<td>9.52</td>
<td>44.76</td>
<td>30.95</td>
<td>5.88</td>
<td>8.70</td>
<td></td>
</tr>
</tbody>
</table>
Table 3: Semantic labeling accuracy
(a) bedroom

<table>
<thead>
<tr>
<th>object type</th>
<th>background</th>
<th>bed</th>
<th>desk</th>
<th>window</th>
<th>mirror</th>
<th>door</th>
<th>nightstand</th>
<th>wardrobe</th>
<th>cabinet</th>
<th>painting</th>
<th>tv</th>
<th>chair</th>
<th>sofa</th>
</tr>
</thead>
<tbody>
<tr>
<td>global (%)</td>
<td>86.90</td>
<td>78.58</td>
<td>29.55</td>
<td>35.58</td>
<td>38.15</td>
<td>19.40</td>
<td>39.66</td>
<td>27.44</td>
<td>0.00</td>
<td>38.70</td>
<td>34.81</td>
<td>9.61</td>
<td>11.10</td>
</tr>
<tr>
<td>local (%)</td>
<td>87.13</td>
<td>80.76</td>
<td>33.10</td>
<td>22.78</td>
<td>42.90</td>
<td>25.47</td>
<td>55.67</td>
<td>25.31</td>
<td>5.46</td>
<td>41.58</td>
<td>32.88</td>
<td>17.20</td>
<td>7.74</td>
</tr>
</tbody>
</table>

(b) living room

<table>
<thead>
<tr>
<th>object type</th>
<th>background</th>
<th>painting</th>
<th>door</th>
<th>cabinet</th>
<th>dining table</th>
<th>window</th>
<th>heater</th>
<th>chair</th>
<th>sofa</th>
<th>coffee table</th>
<th>end table</th>
<th>tv stand</th>
</tr>
</thead>
<tbody>
<tr>
<td>global (%)</td>
<td>91.98</td>
<td>44.66</td>
<td>41.07</td>
<td>7.87</td>
<td>24.24</td>
<td>12.59</td>
<td>0.00</td>
<td>15.46</td>
<td>47.05</td>
<td>42.33</td>
<td>3.87</td>
<td>1.21</td>
</tr>
<tr>
<td>local (%)</td>
<td>93.50</td>
<td>47.50</td>
<td>36.75</td>
<td>16.27</td>
<td>21.80</td>
<td>12.37</td>
<td>11.19</td>
<td>14.95</td>
<td>49.47</td>
<td>42.78</td>
<td>3.99</td>
<td>7.66</td>
</tr>
</tbody>
</table>