Contexts and 3D Scenes

Computer Vision
Jia-Bin Huang, Virginia Tech

Many slides from D. Hoiem
Administrative stuffs

• Final project presentation
  • Dec 1\textsuperscript{st} 3:30 PM – 4:45 PM
  • Goodwin Hall Atrium

• Grading
  • Three instructors; your summary (poster x2)

• Please set up your poster before 3:25 PM
  • Poster boards and easels will be available

• Session 1: 3:35 PM – 4:10 PM
  • Group A present; group B attend the posters

• Session 2: 4:10 PM – 4:45 PM
  • Group B present; group A attend the posters

• Invite your friends!
  • Voting for the Audience Favorite Poster
Context in Recognition

- Objects usually are surrounded by a scene that can provide context in the form of nearby objects, surfaces, scene category, geometry, etc.
Context provides clues for function

- What is this?

These examples from Antonio Torralba
Context provides clues for function

- What is this?

- Now can you tell?
Sometimes context is the major component of recognition

• What is this?
Sometimes context is *the* major component of recognition

• What is this?

• Now can you tell?
More Low-Res

• What are these blobs?
More Low-Res

• The same pixels! (a car)
There are many types of context

- **Local pixels**
  - window, surround, image neighborhood, object boundary/shape, global image statistics

- **2D Scene Gist**
  - global image statistics

- **3D Geometric**
  - 3D scene layout, support surface, surface orientations, occlusions, contact points, etc.

- **Semantic**
  - event/activity depicted, scene category, objects present in the scene and their spatial extents, keywords

- **Photogrammetric**
  - camera height orientation, focal length, lens distortion, radiometric, response function

- **Illumination**
  - sun direction, sky color, cloud cover, shadow contrast, etc.

- **Geographic**
  - GPS location, terrain type, land use category, elevation, population density, etc.

- **Temporal**
  - nearby frames of video, photos taken at similar times, videos of similar scenes, time of capture

- **Cultural**
  - photographer bias, dataset selection bias, visual cliches, etc.

from Divvala et al. CVPR 2009
Cultural context

Jason Salavon: [http://salavon.com/SpecialMoments/Newlyweds.shtml](http://salavon.com/SpecialMoments/Newlyweds.shtml)
Cultural context

“Mildred and Lisa”: Who is Mildred? Who is Lisa?

Andrew Gallagher: http://chenlab.ece.cornell.edu/people/Andy/projectpage_names.html
Cultural context

Age given Appearance

\[
P(f_g|f_a) = \begin{bmatrix} 0.563 \\ 0.437 \end{bmatrix}
\]

![Image of Mildred]

\[
P(f_g|n = \text{Mildred}) = \begin{bmatrix} 0.999 \\ 0.001 \end{bmatrix}
\]

Age given Name

\[
P(f_g|f_a) = \begin{bmatrix} 0.687 \\ 0.313 \end{bmatrix}
\]

![Image of Lisa]

\[
P(f_g|n = \text{Lisa}) = \begin{bmatrix} 0.998 \\ 0.002 \end{bmatrix}
\]
Spatial layout is especially important

1. Context for recognition
Spatial layout is especially important

1. Context for recognition
Spatial layout is especially important

1. Context for recognition
2. Scene understanding
Spatial layout is especially important

1. Context for recognition
2. Scene understanding
3. Many direct applications
   a) Assisted driving
   b) Robot navigation/interaction
   c) 2D to 3D conversion for 3D TV
   d) Object insertion
Spatial Layout: 2D vs. 3D?
Context in Image Space

[Torralba Murphy Freeman 2004]

[Kumar Hebert 2005]

[He Zemel Cerreira-Perpiñán 2004]
But object relations are in 3D...
How to represent scene space?
Wide variety of possible representations

Scene-Level Geometric Description

a) Gist, Spatial Envelope

b) Stages

Figs from Hoiem - Savarese 2011 book
Retinotopic Maps

c) Geometric Context

d) Depth Maps

Figs from Hoiem - Savarese 2011 book
Highly Structured 3D Models

e) Ground Plane
f) Ground Plane with Billboards
g) Ground Plane with Walls

h) Blocks World
i) 3D Box Model

Figs from Hoiem - Savarese 2011 book
Key Trade-offs

• Level of detail: rough “gist”, or detailed point cloud?
  • Precision vs. accuracy
  • Difficulty of inference

• Abstraction: depth at each pixel, or ground planes and walls?
  • What is it for: e.g., metric reconstruction vs. navigation
Low detail, Low/Med abstraction

Holistic Scene Space: “Gist”

Oliva & Torralba 2001

Torralba & Oliva 2002
High detail, Low abstraction

Depth Map

Saxena, Chung & Ng 2005, 2007
Medium detail, High abstraction

Room as a Box

Hedau Hoiem Forsyth 2009
Med-High detail, High abstraction

Complete 3D Layout
Examples of spatial layout estimation

• Surface layout
  • Application to 3D reconstruction

• The room as a box
  • Application to object recognition
Surface Layout: describe 3D surfaces with geometric classes

Sky

Vertical

Support

Non-Planar Porous

Non-Planar Solid

Planar (Left/Center/Right)
The challenge
Our World is Structured

Abstract World

Our World

Image Credit (left): F. Cunin and M.J. Sailor, UCSD
Learn the Structure of the World

Training Images
Infer the most likely interpretation
Geometry estimation as recognition

Region

Features
- Color
- Texture
- Perspective
- Position

Surface Geometry Classifier

Vertical, Planar

Training Data
Use a variety of image cues

Vanishing points, lines

Color, texture, image location

Texture gradient
Surface Layout Algorithm

Input Image → Segmentation → Features → Surface Labels

Features:
- Perspective
- Color
- Texture
- Position

Trained Region Classifier

Training Data

Hoiem Efros Hebert (2007)
Surface Layout Algorithm

Input Image → Multiple Segmentations → Features (Perspective, Color, Texture, Position) → Confidence-Weighted Predictions → Trained Region Classifier → Final Surface Labels

Training Data

Hoiem Efros Hebert (2007)
Surface Description Result
Results

Input Image  Ground Truth  Our Result
Results

Input Image | Ground Truth | Our Result
Results

Input Image

Ground Truth

Our Result
Failures: Reflections, Rare Viewpoint

![Input Image](image1)

![Ground Truth](image2)

![Our Result](image3)

![Input Image](image4)

![Ground Truth](image5)

![Our Result](image6)
Average Accuracy

Main Class: 88%
Subclasses: 61%

<table>
<thead>
<tr>
<th>Main Class</th>
<th>Support</th>
<th>Vertical</th>
<th>Sky</th>
</tr>
</thead>
<tbody>
<tr>
<td>Support</td>
<td>0.84</td>
<td>0.15</td>
<td>0.00</td>
</tr>
<tr>
<td>Vertical</td>
<td>0.09</td>
<td>0.90</td>
<td>0.02</td>
</tr>
<tr>
<td>Sky</td>
<td>0.00</td>
<td>0.10</td>
<td>0.90</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Vertical Subclass</th>
<th>Left</th>
<th>Center</th>
<th>Right</th>
<th>Porous</th>
<th>Solid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left</td>
<td>0.37</td>
<td>0.32</td>
<td>0.08</td>
<td>0.09</td>
<td>0.13</td>
</tr>
<tr>
<td>Center</td>
<td>0.05</td>
<td>0.56</td>
<td>0.12</td>
<td>0.16</td>
<td>0.12</td>
</tr>
<tr>
<td>Right</td>
<td>0.02</td>
<td>0.28</td>
<td>0.47</td>
<td>0.13</td>
<td>0.10</td>
</tr>
<tr>
<td>Porous</td>
<td>0.01</td>
<td>0.07</td>
<td>0.03</td>
<td>0.84</td>
<td>0.06</td>
</tr>
<tr>
<td>Solid</td>
<td>0.04</td>
<td>0.20</td>
<td>0.04</td>
<td>0.17</td>
<td>0.55</td>
</tr>
</tbody>
</table>
Automatic Photo Popup

Labeled Image

Fit Ground-Vertical Boundary with Line Segments

Form Segments into Polylines

Cut and Fold

Final Pop-up Model

[Hoiem Efros Hebert 2005]
Automatic Photo Popup
Mini-conclusions

• Can learn to predict surface geometry from a single image
Interpretation of indoor scenes
Vision = assigning labels to pixels?
Vision = interpreting within physical space
Physical space needed for affordance

Is this a good place to sit?

Could I stand over here?

Can I put my cup here?

Walkable path
Physical space needed for recognition

Apparent shape depends strongly on viewpoint
Physical space needed for recognition
Physical space needed to predict appearance
Physical space needed to predict appearance
Key challenges

• How to represent the physical space?
  • Requires seeing beyond the visible

• How to estimate the physical space?
  • Requires simplified models
  • Requires learning from examples
Our Box Layout

• Room is an oriented 3D box
  • Three vanishing points specify orientation
  • Two pairs of sampled rays specify position/size
Our Box Layout

• Room is an oriented 3D box
  • Three vanishing points (VPs) specify orientation
  • Two pairs of sampled rays specify position/size

Another box consistent with the same vanishing points
Image Cues for Box Layout

• Straight edges
  • Edges on floor/wall surfaces are usually oriented towards VPs
  • Edges on objects might mislead

• Appearance of visible surfaces
  • Floor, wall, ceiling, object labels should be consistent with box
Box Layout Algorithm

1. Detect edges

2. Estimate 3 orthogonal vanishing points

3. Apply region classifier to label pixels with visible surfaces
   • Boosted decision trees on region based on color, texture, edges, position

4. Generate box candidates by sampling pairs of rays from VPs

5. Score each box based on edges and pixel labels
   • Learn score via structured learning

6. Jointly refine box layout and pixel labels to get final estimate
Evaluation

• Dataset: 308 indoor images
  • Train with 204 images, test with 104 images
Experimental results

Detected Edges  
Surface Labels  
Box Layout

Detected Edges  
Surface Labels  
Box Layout
Experimental results

Detected Edges

Surface Labels

Box Layout

Detected Edges

Surface Labels

Box Layout
Experimental results

• Joint reasoning of surface label / box layout helps
  • Pixel error: 26.5% → 21.2%
  • Corner error: 7.4% → 6.3%

• Similar performance for cluttered and uncluttered rooms
Mini-Conclusions

• Can fit a 3D box to the rooms boundaries from one image
  • Robust to occluding objects
  • Decent accuracy, but still much room for improvement
Using room layout to improve object detection

Box layout helps

1. Predict the appearance of objects, because they are often aligned with the room
2. Predict the position and size of objects, due to physical constraints and size consistency

2D Bed Detection
3D Bed Detection with Scene Geometry

Hedau, Hoiem, Forsyth, ECCV 2010, CVPR 2012
Search for objects in room coordinates

Recover Room Coordinates

Rectify Features to Room Coordinates

Rectified Sliding Windows

Hedau Forsyth Hoiem (2010)
Reason about 3D room and bed space

Joint Inference with Priors
• Beds close to walls
• Beds within room
• Consistent bed/wall size
• Two objects cannot occupy the same space

Hedau Forsyth Hoiem (2010)
3D Bed Detection from an Image

True positives

False positives
Generic boxy object detection
Generic boxy object detection
Generic boxy object detection
Mini-Conclusions

• Simple room box layout helps detect objects by predicting appearance and constraining position

• We can search for objects in 3D space and directly evaluate on 3D localization
Predicting complete models from RGBD

Key idea: create **complete** 3D scene hypothesis that is **consistent** with observed depth and appearance.
Overview of approach

RGB-D Input

Object Proposals

Exemplar Region Retrieval

Retrieved Region

Source 3D Model

3D Model Fitting

Transferred Model

Transferred Model

Annotated Scene

Composing

Retrieved Region

Source 3D Model

Layout Proposals
Example result (fully automatic)
Things to remember

• Objects should be interpreted in the context of the surrounding scene
  • Many types of context to consider

• Spatial layout is an important part of scene interpretation, but many open problems
  • How to represent space?
  • How to learn and infer spatial models?
  • Important to see beyond the visible

• Consider trade-off of abstraction vs. precision