Context and Spatial Layout

Slides mostly from
Derek Hoiem and Svetlana Lazebnik
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
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

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

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

\[ P(f_g \mid n = \text{Lisa}) = \begin{bmatrix} 0.998 \\ 0.002 \end{bmatrix} \]

Age given Name

Andrew Gallagher: [http://chenlab.ece.cornell.edu/people/Andy/projectpage_names.html](http://chenlab.ece.cornell.edu/people/Andy/projectpage_names.html)
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?
How to represent scene space?
Wide variety of possible representations

Scene-Level Geometric Description

a) Gist, Spatial Envelope

b) Stages

Figs from Hoiem/Savarese Draft
Retinotopic Maps

c) Geometric Context

d) Depth Maps

Figs from Hoiem/Savarese Draft
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 Draft
Key Trade-offs

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

• Abstraction: depth at each pixel, or ground planes and walls?
  – What is it for: e.g., metric reconstruction vs. navigation
Examples of spatial layout estimation

• GIST

• Surface layout
  – Application to 3D reconstruction

• The room as a box
  – Application to object recognition
Wide variety of possible representations

Scene-Level Geometric Description

a) Gist, Spatial Envelope

b) Stages

Figs from Hoiem/Savarese Draft
Retinotopic Maps

c) Geometric Context

d) Depth Maps

Figs from Hoiem/Savarese Draft
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 Draft
Low detail, Low/Med abstraction

Holistic Scene Space: “Gist”

Idea:
Partition the image to anchor the structure of the subimages to their location in the image.
GIST

Split 

Fourier transformation 

Multiscale oriented filters 

$n \times n \times k$ - vector 

Gist-descriptor 

$n \times n = \# \text{ partitions} 

k = \# \text{scales} \times \# \text{orientations}$
GIST
GIST
GIST

$n \times n \times k = 4 \times 4 \times 5$ frequencies $\times 6$ orientations

$= 480$-dim. Vector
GIST
GIST for image retrieval
GIST for image retrieval
Use of GIST

Database of images with GIST computed

Test image
Use of GIST

Database of images with GIST computed

Test image

Retrieve similar images
Use of GIST

Database of images with GIST computed

Retrieve similar images and transfer scene information…

Test image
Examples of spatial layout estimation

• GIST

• 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
Automatic Photo Popup

D. Hoiem  A.A. Efros  M. Hebert
Carnegie Mellon University
Geometry estimation as recognition

Region → Features → Surface Geometry Classifier → Vertical, Planar

Features: Color, Texture, Perspective, Position

Training Data

...
Use a variety of image cues

Vanishing points, lines

Color, texture, image location

Texture gradient
Use a variety of image cues

- Material
- Image Location
- 3D Geometry

<table>
<thead>
<tr>
<th>Feature Descriptions</th>
<th>Num</th>
<th>Used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Color</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>C1. RGB values: mean</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>C2. HSV values: conversion from mean RGB values</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>C3. Hue: histogram (5 bins) and entropy</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>C4. Saturation: histogram (3 bins) and entropy</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Texture</td>
<td>29</td>
<td>13</td>
</tr>
<tr>
<td>T1. DOOG Filters: mean abs response</td>
<td>12</td>
<td>7</td>
</tr>
<tr>
<td>T2. DOOG Filters: mean of variables in T1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>T3. DOOG Filters: id of max of variables in T1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>T4. DOOG Filters: (max - median) of variables in T1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>T5. Textons: mean abs response</td>
<td>12</td>
<td>7</td>
</tr>
<tr>
<td>T6. Textons: max of variables in T5</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>T7. Textons: (max - median) of variables in T5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Location and Shape</td>
<td>12</td>
<td>10</td>
</tr>
<tr>
<td>L1. Location: normalized x and y, mean</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>L2. Location: norm. x and y, 10th and 90th percentile</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>L3. Location: norm. y wrt horizon, 10th and 90th percentile</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>L4. Shape: number of superpixels in constellation</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>L5. Shape: number of sides of convex hull</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>L6. Shape: num pixels/area(voxel hull)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>L7. Shape: whether the constellation region is contiguous</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3D Geometry</td>
<td>35</td>
<td>28</td>
</tr>
<tr>
<td>G1. Long Lines: total number in constellation region</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>G2. Long Lines: % of nearly parallel pairs of lines</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>G3. Line Intersection: hist. over 12 orientations, entropy</td>
<td>13</td>
<td>11</td>
</tr>
<tr>
<td>G4. Line Intersection: % right of center</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>G5. Line Intersection: % above center</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>G6. Line Intersection: % far from center at 8 orientations</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>G7. Line Intersection: % very far from center at 8 orientations</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>G8. Texture gradient: x and y “edginess” (T2) center</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>
Surface Layout Algorithm

Input Image → Segmentation → Features (Perspective, Color, Texture, Position) → Surface Labels

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 | Ground Truth | Our Result
Average Accuracy

Main Class: 88%
Subclasses: 61%

### Main Class

<table>
<thead>
<tr>
<th></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>

### Vertical Subclass

<table>
<thead>
<tr>
<th></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

Input Images

Automatic Photo Pop-up
Automatic Photo Popup

Input Image

[Liebowitz et al. 1999]

Automatic Photo Pop-up (30 sec)!
Failures

Foreground Objects
Mini-conclusions

• Can learn to predict surface geometry from a single image
• Very rough models, much room for improvement
Examples of spatial layout estimation

• GIST

• Surface layout
  – Application to 3D reconstruction

• The room as a box
  – Application to object recognition
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?

Walkable path

Could I stand over here?

Can I put my cup here?
Physical space needed for recognition

Apparent shape depends strongly on viewpoint
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

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
3D Bed Detection from an Image

True positives

False positives
Generic boxy object detection
Generic boxy object detection
Generic boxy object detection
Good localization in image doesn’t mean good localization in 3D
Refining 3D location

- Refit bounding box by detecting bottom edges of objects and furniture legs
3D Evaluation

Ground Truth

Estimate

Ground Truth

Estimate
Mini-Conclusions

- Our 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.
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