Lecture 16

3D scene understanding
Announcements

• No class on Wed (ECCV deadline!)

• In-class presentations next week, two parallel two-tracks sessions
  • 12:30pm - 2:30pm, March 19
  • Room 1: Oshman 125
  • Room 2: 450 Serra Mall, 300-300

• 3.5 minutes for each presentation including Q&A

• It’s a team presentation

• In-class or Piazza questions count toward attendance evaluation
Announcements

• See Piazza and website for more information on the format for presentation and write up
• CA session on Friday to discuss expectations for final report
• Thanks for the online evaluations!
• Your feedback is extremely important!
What does it mean to understand a scene?
Computers can reconstruct 3D spaces better than humans!

Snively et al., 06-08
Computers can reconstruct 3D spaces better than humans!
Courtesy of Luminar
It this useful?
It is about “where” things are...
Is this sufficient?
Machine vision is...

...not just about “where”,
also about “what”
Image-to-labels paradigm

image

labels
Computers can recognize objects better than humans! (*)

Building

Person

Strawberry

Person

Strawberry

Road
Building facade

Road
Objects are constrained by the 3D space

The 3D space is shaped by its objects

Modeling this interplay is critical for 3D perception!
Humans perceive the world in 3D

Biederman, Mezzanotte and Rabinowitz, 1982
Visual processing in the brain

where pathway
(dorsal stream)

what pathway
(ventral stream)

V1
Visual processing in the brain

V1

where pathway (dorsal stream)

what pathway (ventral stream)

Pre-frontal cortex
Two sides of one coin

3D Reconstruction
- 3D shape recovery
- 3D scene reconstruction
- Camera localization
- Pose estimation

2D Recognition
- Object detection
- Texture classification
- Target tracking
- Activity recognition
Two sides of one coin

3D Reconstruction
- 3D shape recovery
- 3D scene reconstruction
- Camera localization
- Pose estimation

2D Recognition
- Object detection
- Texture classification
- Target tracking
- Activity recognition

Perceiving the World in 3D
- Modeling objects and their 3D properties
- Modeling interaction among objects and space
- Modeling relationships of object/space across views
Outline

• Modeling objects and their 3D properties
• Modeling interaction among objects and space
• Modeling relationships of objects across views
Detecting objects and estimating their 3D properties
Results

CAR  a=330  e=15  d=7

MOUSE  a=300  e=45  d=23

CAR  a=150  e=15  d=7

SHOE  a=240  e=45  d=11

3D object dataset [Savarese & Fei-Fei 07]
Results

CHAIR  a=0  e=30  d=7

BED    a=30  e=15  d=2.5

TABLE  a=60  e=15  d=2

SOFA   a=345 e=15  d=3.5
       a=60  e-30  d=2.5

ImageNet dataset [Deng et al. 2010]
Results
Examples of failure (wrong category)

This can’t be a shoe!
Outline

• Modeling objects and their 3D properties
• Modeling interaction among objects and space
• Modeling relationships of objects across views
Scene understanding is an interplay between objects and space
3D space is shaped by its objects
Objects are placed into 3D space
A first attempt....

Interactions object-ground
A first attempt....

Hoiem et al, 2006-2008
Bao & Savarese, 2008-2012
A first attempt....

Hoiem et al, 2006-2008
Bao & Savarese, 2008-2012
Generalization #1

Interactions between:
- Objects-space
- Object-object

Oliva & Torralba, 2007
Rabinovich et al, 2007
Li & Fei-Fei, 2007
Vogel & Schiele, 2007
Desai et al, 2009
Sadeghi & Farhadi, 2011
Li et al, 2012
Hoiem et al, 2006
Herdau et al., 2009
Gupta et al, 2010
Fouhey et al, 2012
3D geometric phrases

Choi, Chao, Pantofaru, Savarese, CVPR 13, IJCV 15
3D Geometric Phrases

Choi, Chao, Pantofaru, Savarese, CVPR 13, IJCV 15

- w/o annotations
- Compact
- View-invariant
Enable contextual awareness!

Sofa, Coffee Table, Chair, Bed, Dining Table, Side Table

Estimated Layout

3D Geometric Phrases
Results: Object Detection

Average Precision %

<table>
<thead>
<tr>
<th>Object</th>
<th>Felzenszwalb et al.</th>
<th>3DGP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sofa</td>
<td>20</td>
<td>30</td>
</tr>
<tr>
<td>Table</td>
<td>20</td>
<td>30</td>
</tr>
<tr>
<td>Chair</td>
<td>20</td>
<td>30</td>
</tr>
<tr>
<td>Bed</td>
<td>56.9</td>
<td>65.2</td>
</tr>
<tr>
<td>D. Table</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S. Table</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>65.2</td>
<td>65.2</td>
</tr>
</tbody>
</table>

Indoor scene dataset [Choi et al., 12]
Outline

• Modeling objects and their 3D properties
• Modeling interaction among objects and space
• Modeling relationships of objects across views
Modeling relationships of objects across views

- Interaction between object-space
- Interaction among objects
- Transfer semantics across views
Modeling relationships of objects across views

- Interaction between object-space
- Interaction among objects
- Transfer semantics across views
Semantic structure from motion

• Measurements I
  • Points (x, y, scale)
  • Objects (x, y, scale, pose)
  • Regions (x, y, pose)

• Model Parameters:
  • Q = 3D points
  • O = 3D objects
  • B = 3D regions
  • C = cam. prm. K, R, T
Semantic structure from motion

\[ \{Q, O, B, C\} = \arg \max_{Q,O,B,C} \Psi(Q, O, B, C; I) \]

- **Measurements I**
  - Points \((x, y, \text{scale})\)
  - Objects \((x, y, \text{scale, pose})\)
  - Regions \((x, y, \text{pose})\)

- **Model Parameters:**
  - \(Q = 3D\) points
  - \(O = 3D\) objects
  - \(B = 3D\) regions
  - \(C = \text{cam. prm. } K, R, T\)
Semantic structure from motion

\[
\{Q, O, B, C\} = \arg \max_{Q, O, B, C} \prod_s \psi^{CQ}_s \prod_t \psi^{CO}_t \prod_r \psi^{CB}_r
\]

- **Measurements**
  - Points \((x, y, \text{scale})\)
  - Objects \((x, y, \text{scale, pose})\)
  - Regions \((x, y, \text{pose})\)

- **Model Parameters**
  - \(Q = 3D\) points
  - \(O = 3D\) objects
  - \(B = 3D\) regions
  - \(C = \text{cam. prm. } K, R, T\)
SSFM: point-level compatibility

\[ \{Q, O, B, C\} = \arg \max_{Q,O,B,C} \prod_s \psi^{CQ}_s \prod_t \psi^{CO}_t \prod_r \psi^{CB}_r \]

- **Measurements**
  - Points \((x, y, \text{scale})\)
  - Objects \((x, y, \text{scale, pose})\)
  - Regions \((x, y, \text{pose})\)

- **Model Parameters**
  - \(Q = 3D\) points
  - \(O = 3D\) objects
  - \(B = 3D\) regions
  - \(C = \text{cam. prm. } K, R, T\)
SSFM: point-level compatibility

\[
\{Q, \mathcal{O}, B, C\} = \arg \max_{Q, \mathcal{O}, B, C} \prod_s \psi_{s}^{CQ} \prod_t \psi_{t}^{CO} \prod_r \psi_{r}^{CB}
\]

- **Measurements I**
  - Points \((x, y, \text{scale})\)
  - Objects \((x, y, \text{scale}, \text{pose})\)
  - Regions \((x, y, \text{pose})\)

- **Model Parameters:**
  - \(Q = 3D\) points
  - \(O = 3D\) objects
  - \(B = 3D\) regions
  - \(C = \text{cam. prm.} \ K, R, T\)

Point re-projection error

\[
\prod_s \psi_{s}^{CQ} \propto \prod_{i} \prod_{k} \exp\left(-\frac{(q_{i}^{k} - q_{u_{i}}^{k})^2}{\sigma_{q}}\right)
\]
SSFM: Object-level compatibility

\[
\{Q, O, B, C\} = \arg \max_{Q, O, B, C} \prod_s \Psi_s^{CQ} \prod_t \Psi_t^{CO} \prod_r \Psi_r^{CB}
\]

- **Measurements I**
  - Points \((x, y, \text{scale})\)
  - Objects \((x, y, \text{scale, pose})\)
  - Regions \((x, y, \text{pose})\)

- **Model Parameters:**
  - \(Q = 3D \text{ points}\)
  - \(O = 3D \text{ objects}\)
  - \(B = 3D \text{ regions}\)
  - \(C = \text{cam. prm. } K, R, T\)
SSFM: Object-level compatibility

\[
\{Q, O, B, C\} = \text{arg} \max_{Q,O,B,C} \prod_s \Psi_{s}^{CQ} \prod_t \Psi_{t}^{CO} \prod_r \Psi_{r}^{CB}
\]

• Measurements I
  • Points (x,y, scale)
  • Objects (x,y, scale, pose)
  • Regions (x,y, pose)

• Model Parameters:
  • Q = 3D points
  • O = 3D objects
  • B = 3D regions
  • C = cam. prm. K, R, T

Object “re-projection” error

\[
\Psi_{t}^{CO} \propto \prod_{t}^{N_t} \prod_{k}^{N_k} (1 - \text{Pr}(o|O_t, C^k)))
\]
SSFM: Object-level compatibility

- Agreement with measurements is computed using position, pose and scale
SSFM: Object-level compatibility

- Agreement with measurements is computed using position, pose and scale
SSFM with interactions

\( \{Q, O, B, C\} = \arg \max_{Q, O, B, C} \prod_s \psi_{sQ} \prod_t \psi_{tO} \prod_r \psi_{rB} \prod_{t,s} \psi_{t,sQ} \prod_{t,r} \psi_{t,rO} \prod_{r,s} \psi_{r,sB} \)

- **Measurements I**
  - Points \((x,y,\text{scale})\)
  - Objects \((x,y,\text{scale, pose})\)
  - Regions \((x,y,\text{pose})\)

- **Model Parameters:**
  - \(Q = 3D\) points
  - \(O = 3D\) objects
  - \(B = 3D\) regions
  - \(C = \text{cam. prm. } K, R, T\)

- Interactions of points, regions and objects across views
- Interactions among object-regions-points

Bao, Bagra, Chao, Savarese
CVPR 2012
SSFM with interactions

\[ \{Q, O, B, C\} = \arg \max_{Q,O,B,C} \prod_s \psi_{sQ} \prod_t \psi_{tO} \prod_r \psi_{rB} \prod_{t,s} \psi_{t,sQ} \prod_{t,r} \psi_{t,rO} \prod_{r,s} \psi_{r,sB} \]

Object-Region Interactions:

- Measurements I
  - Points \((x, y, \text{scale})\)
  - Objects \((x, y, \text{scale}, \text{pose})\)
  - Regions \((x, y, \text{pose})\)

- Model Parameters:
  - \(Q = 3D\) points
  - \(O = 3D\) objects
  - \(B = 3D\) regions
  - \(C = \text{cam. prm. } K, R, T\)
Object-Region Interactions:

- Measurements I:
  - Points \( (x, y, \text{scale}) \)
  - Objects \( (x, y, \text{scale, pose}) \)
  - Regions \( (x, y, \text{pose}) \)

- Model Parameters:
  - \( Q = \) 3D points
  - \( O = \) 3D objects
  - \( B = \) 3D regions
  - \( C = \) cam. prm. \( K, R, T \)

\[
\{Q, O, B, C\} = \underset{Q, O, B, C}{\text{arg max}} \prod_s \psi_s^{CQ} \prod_t \psi_t^{CO} \prod_r \psi_r^{CB} \prod_{t,s} \psi_{t,s}^{OQ} \prod_{t,r} \psi_{t,r}^{OB} \prod_{r,s} \psi_{r,s}^{BQ}
\]
SSFM with interactions

\begin{align*}
\{Q, O, B, C\} &= \arg \max_{Q, O, B, C} \prod_s \psi^{CQ}_s \prod_t \psi^{CO}_t \prod_r \psi^{CB}_r \\
& \quad \times \prod_{t,s} \psi^{OQ}_{t,s} \prod_{t,r} \psi^{OB}_{t,r} \prod_{r,s} \psi^{BQ}_{r,s}
\end{align*}

Object-point Interactions:

- Measurements:
  - Points \((x, y, \text{scale})\)
  - Objects \((x, y, \text{scale}, \text{pose})\)
  - Regions \((x, y, \text{pose})\)

- Model Parameters:
  - \(Q\) = 3D points
  - \(O\) = 3D objects
  - \(B\) = 3D regions
  - \(C\) = cam. prm. \(K, R, T\)
SSFM with interactions

\[ \{Q, O, B, C\} = \arg \max_{Q, O, B, C} \prod_s \psi_s^{CQ} \prod_t \psi_t^{CO} \prod_r \psi_r^{CB} \prod_{t,s} \psi_{t,s}^{OQ} \prod_{t,r} \psi_{t,r}^{OB} \prod_{r,s} \psi_{r,s}^{BQ} \]

Object-point Interactions:

- Measurements I
  - Points \((x, y, \text{scale})\)
  - Objects \((x, y, \text{scale, pose})\)
  - Regions \((x, y, \text{pose})\)

Model Parameters:
- \(Q = 3D\) points
- \(O = 3D\) objects
- \(B = 3D\) regions
- \(C = \text{cam. prm. } K, R, T\)
Solving the SSFM problem

\[ \{Q, O, B, C\} = \arg \max_{Q, O, B, C} \Psi(Q, O, B, C; I) \]

- Modified Reversible Jump Markov Chain Monte Carlo (RJ-MCMC) sampling algorithm [Dellaert et al., 2000]

- Initialization of the cameras, objects, and points are critical for the sampling

- Initialize configuration of cameras using:
  - SFM
  - consistency of object/region properties across views
Results

Input images

- Wide baseline
- Background clutter
- Limited visibility
- Un-calibrated cameras
Results

Input images
Input images

Results

- Car
- Person
- Tree
- Sky
- Street
- Building
- Else
Results

Input images

From the office dataset [Bao et al., 11]
Results

Input images

From the office dataset [Bao et al., 11]
Results

Average precision in localizing objects in the 3D space

<table>
<thead>
<tr>
<th></th>
<th>Hoiem et al. 2011</th>
<th>SSFM no int.</th>
<th>SSFM</th>
</tr>
</thead>
<tbody>
<tr>
<td>FORD CAMPUS</td>
<td>21.4%</td>
<td>32.7%</td>
<td>43.1%</td>
</tr>
<tr>
<td>OFFICE</td>
<td>15.5%</td>
<td>20.2%</td>
<td>21.6%</td>
</tr>
</tbody>
</table>

Average precision in detecting objects in the 2D image

<table>
<thead>
<tr>
<th>DPM [1]</th>
<th>SSFM 2 views no int.</th>
<th>SSFM 2 views</th>
<th>SSFM 4 views</th>
</tr>
</thead>
<tbody>
<tr>
<td>54.5%</td>
<td>61.3%</td>
<td>62.8%</td>
<td><strong>66.5%</strong></td>
</tr>
</tbody>
</table>

FORD CAMPUS dataset [Pandey et al., 09]  
Office dataset [Bao et al., 11]

## Results

<table>
<thead>
<tr>
<th></th>
<th>Camera translation error</th>
<th>Camera rotation error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>SFM</strong></td>
<td><strong>SSFM</strong></td>
</tr>
<tr>
<td><strong>FORD CAMPUS</strong></td>
<td>26.5°</td>
<td>19.9°</td>
</tr>
<tr>
<td><strong>OFFICE</strong></td>
<td>8.5°</td>
<td>4.7°</td>
</tr>
<tr>
<td><strong>STREET</strong></td>
<td>27.1°</td>
<td>17.6°</td>
</tr>
</tbody>
</table>

FORD CAMPUS dataset [Pandey et al., 09]
Office dataset [Bao et al., 11]
Street dataset [Bao et al., 11]
Wide-baseline feature correspondence
SSFM Source code available!
Please visit: http://www.eecs.umich.edu/vision/research.html
Toward large scale scene understanding
Large-scale scene parsing

Armeni, Sener, Zamir, Jiang, Brilakis, Fischer, Savarese, 2016
Large-scale scene parsing

Armeni, Sener, Zamir, Jiang, Brilakis, Fischer, Savarese, 2016
Toward fine-grained 3D scene parsing

Tchapmi et al., 3DV, 2017
Stanford Large-Scale Indoor dataset

Armeni, Sener, Zamir, Jiang, Brilakis, Fischer, Savarese, 2016

6 buildings
~500 rooms
~6000m² area
~6000 Building Elements
Building information management

http://buildingparser.stanford.edu
Applications

Robotic navigation

Objects understanding

Space understanding

Scene understanding

Construction monitoring

Mobile vision

Safe driving
Visual intelligence and large scale information management

Golparvar-Fard, Pena-Mora, Savarese, 2008-2012
James R. Croes Medal, October 2013 (from the American Society of Civil of Engineers)

Automatic coordination of construction progress can lead to huge savings (10 billions USD/year) in construction business!

[Census Bureau, www.census.gov, 2007]
Images are cheap!
Our revolution
Our revolution
<table>
<thead>
<tr>
<th>Task</th>
<th>Start Date</th>
<th>End Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concrete Excavate Upper Footings Area C</td>
<td>15SEP06</td>
<td>14NOV06</td>
</tr>
<tr>
<td>Concrete Excavate Footings Area D</td>
<td>25SEP06</td>
<td>27SEP06</td>
</tr>
<tr>
<td>Concrete Pour Footings Area D</td>
<td>30SEP06</td>
<td>16NOV06</td>
</tr>
<tr>
<td>Concrete Pour Upper Footings Area D</td>
<td>02OCT06</td>
<td>03OCT06</td>
</tr>
<tr>
<td>Concrete Excavate Column Pads Area A</td>
<td>11OCT06</td>
<td>14NOV06</td>
</tr>
<tr>
<td>Concrete Pour Column Pads Area A</td>
<td>10OCT06</td>
<td>14NOV06</td>
</tr>
<tr>
<td>Concrete Waterproof First Lift for Drain Tile</td>
<td>10OCT06</td>
<td>06FEB07</td>
</tr>
<tr>
<td>Concrete Form/Pour Upper Walls Area C</td>
<td>11OCT06</td>
<td>16NOV06</td>
</tr>
<tr>
<td>Exterior Perimeter Drain Area A</td>
<td>11OCT06</td>
<td>17NOV06</td>
</tr>
<tr>
<td>Concrete Pour Walls Area B</td>
<td>19OCT06</td>
<td>17NOV06</td>
</tr>
<tr>
<td>Concrete Forms Walls Area D</td>
<td>08NOV06</td>
<td>29NOV06</td>
</tr>
<tr>
<td>Concrete Mock-Up</td>
<td>09NOV06</td>
<td>08NOV06</td>
</tr>
</tbody>
</table>

**Timeline:**
- **12/02/2006; 1:13:00 PM (As-built):**
  - Ahead of Schedule
  - On Schedule
- **12/02/2006; 1:13:00 PM (As-planned):**
  - Behind Schedule
A new experimental platform: The JackRabbbot
A new experimental platform: The JackRabbit

• Sensors
  – Planar laser scanner
  – 3D stereo vision/laser scanner
  – Cameras
  – IMU+GPS
A new experimental platform: The JackRabbit
Hope you enjoyed this course

Good luck on your presentations next week!