Visual Understanding without Naming: Bypassing the “Language Bottleneck”

Alexei (Alyosha) Efros
UC Berkeley
Collaborators

Abhinav Gupta
Scott Satkin
David Fouhey
Martial Hebert
Natasha Kholgade
Yaser Sheikh

Vincent Delaitre
Mathieu Aubry
Bryan Russell
Ivan Laptev
Josef Sivic

Jun-Yan Zhu
Yong Jae Lee
What do we mean by Visual Understanding?
Object naming -> Object categorization

sky
building
flag
banner
face
street lamp
wall
bus
cars

slide by Fei Fei, Fergus & Torralba
• Not one-to-one:
  – Much is unnamed
• Not one-to-one:
  - Much is unnamed
Verbs (actions)

sitting
Visual “sitting”
The Language Bottleneck

Visual World

words

Scene understanding, spatial reasoning, prediction, image retrieval, image synthesis, etc.
1. 3D Human Affordances
2. 3D Object Correspondence
3. User-in-the-visual-loop

Scene understanding, spatial reasoning, prediction, image retrieval, image synthesis, etc.
From 3D Scene Geometry to Human Workspaces

Abhinav Gupta, Scott Satkin, Alexei Efros and Martial Hebert
CVPR’11
Object Naming

Couch

Lamp

Table

Couch
Is there a **couch** in the image?
Where can I sit?
3D Indoor Image Understanding

Spatial Layout

Objects

Hoiem et al. IJCV’07, Delage et al. CVPR’06, Hedau et al. ICCV’09., Lee et al. NIPS’10, Wang et al. ECCV’10
Human Centric Scene Understanding

Reasoning in terms of set of allowable actions

Can Sit
Can Move
Can Push
Can Walk
Sitting
Pose-defined Vocabulary

Sitting

Poses

Motion Capture
Human Workspace

3D Scene Geometry

Joint Space of Human-Scene Interactions
Qualitative Representation
3D Scene Geometry

• Each scene modeled by
  • Layout of the Room
  • Layout of the Objects

• Room Represented by inside-out box

• Objects represented by occupied voxels.

References:
Hedau et al. ICCV’09., Lee et al. NIPS’10, Wang et al. ECCV’10
Goal

Where would the Human Block fit?
Free Space Constraint: No Intersection between Human Block and Objects
Human Scene Interactions

Support Constraint: Presence of Objects for Interaction
Ground-Truth
3D Geometry

Data Source:
Google 3D Warehouse
Ground-Truth
3D Geometry

Data Source:
Google 3D Warehouse
Extracting 3D Geometry

• Estimating 3D Scene Geometry from a single image is an extremely difficult problem.

• Build on work in 3D Scene Understanding of [Hedau’09] and [Lee’10]
Subjective Scene Interpretation
Summary
The Inverse Problem
People Watching: Human Actions as a Cue for Single-View Geometry

David Fouhey, Vincent Delaitre, Abhinav Gupta, Alexei Efros, Ivan Laptev, Josef Sivic
ECCV 2012
Humans as Active Sensors

Input: Timelapse

Output: 3D Understanding
Our Approach

Timelapse

Pose Detections
Detecting Human Actions

Standing  |  Sitting  |  Reaching

Yang and Ramanan ‘11
Train Separate Detectors for Each Pose
Our Approach

Timelapse → Pose Detections

Estimate Functional Regions from Poses
From Poses to Functional Regions

Sittable Regions at Pelvic Joint
From Poses to Functional Regions

Walkable Regions at Feet
Affordance Constraints

Reachable Regions at Hands
Our Approach

Timelapse → Pose Detections → Functional Regions

3D Room Hypotheses From Appearance
Our Approach

Timelapse → Pose Detections → Functional Regions

Score 3D Room Hypotheses With Appearances + Affordances
Our Approach

Timelapse → Pose Detections → Functional Regions

Estimate Free-Space
Results
Qualitative Example
Qualitative Example

Original video
40 Timelapse videos from Youtube Evaluated on room layout estimation.

<table>
<thead>
<tr>
<th>Location</th>
<th>Appearance Only</th>
<th>People Only</th>
<th>Appearance + People</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lee et al. '09</td>
<td>64.1%</td>
<td>70.4%</td>
<td></td>
</tr>
<tr>
<td>Hedau et al. '09</td>
<td></td>
<td>74.9%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>70.8%</td>
<td><strong>82.5%</strong></td>
</tr>
</tbody>
</table>

Does equivalently or better 93% of the time
Seeing 3D chairs:
Exemplar part-based 2D-3D alignment using a large dataset of CAD models
CVPR 2014

Mathieu Aubry (INRIA)
Daniel Maturana (CMU)
Alexei Efros (UC Berkeley)
Bryan Russell (Intel)
Josef Sivic (INRIA)
Sit on the chair!
Classification

Ex: ImageNet Challenge, Pascal VOC classification.
Detection

Ex: Pascal VOC detection.
Segmentation

Ex: Pascal VOC segmentation.
Our goal
1980s: 2D-3D Instance Alignment

[Lowe AI 1987]

[Huttenlocher and Ullman IJCV 1990]

[Faugeras&Hebert’86], [Grimson&Lozano-Perez’86], ...
Recent: **3D category recognition**

3D DPMs: [Herjati & Ramanan’12], [Pepik et al.12], [Zia et al.’13], ...  
Simplified part models: [Xiang & Savarese’12], [Del Pero et al.’13]

Cuboids: [Xiao et al.’12] [Fidler et al.’12]  
Blocks world revisited: [Gupta et al.’12]

See also: [Glasner et al.’11], [Fouhey et al.’13], [Satkin & Hebert’13], [Choi et al. ‘13], [Hejrati and Ramanan ‘14], [Savarese and Fei-Fei ‘07]...
Approach: data-driven

1394 3D models from internet
Difficulty: viewpoint
Approach: use 3D models

62 views
Style

Viewpoint
Difficulty: approximate style
Difficulty: approximate style
Difficulty: approximate style
Approach: part-based model
Approach overview

3D collection

Render views

Select parts

Match CG->real image

Select the best matches
Select discriminative parts
How to select discriminative parts?

Best exemplar-LDA classifiers

Approach: CG-to-photograph

Implementation: exemplar-LDA
How to compare matches?

<table>
<thead>
<tr>
<th>Patches</th>
<th>Detectors</th>
<th>Matches</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Chair" /></td>
<td>((w_1, b_1))</td>
<td><img src="image2.png" alt="Matches" /></td>
</tr>
<tr>
<td><img src="image3.png" alt="Chair" /></td>
<td>((w_2, b_2))</td>
<td></td>
</tr>
<tr>
<td><img src="image4.png" alt="Chair" /></td>
<td>((w_3, b_3))</td>
<td></td>
</tr>
</tbody>
</table>
### How to compare matches?

<table>
<thead>
<tr>
<th>Patches</th>
<th>Detectors</th>
<th>Affine Calibration with negative data</th>
<th>Matches</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Chair" /></td>
<td>$(w_1, b_1)$</td>
<td>$(a_1 w_1, b'_1)$</td>
<td><img src="image2" alt="Match" /></td>
</tr>
<tr>
<td><img src="image1" alt="Chair" /></td>
<td>$(w_2, b_2)$</td>
<td>$(a_2 w_2, b'_2)$</td>
<td><img src="image3" alt="Match" /></td>
</tr>
<tr>
<td><img src="image1" alt="Chair" /></td>
<td>$(w_3, b_3)$</td>
<td>$(a_3 w_3, b'_3)$</td>
<td><img src="image4" alt="Match" /></td>
</tr>
</tbody>
</table>

See paper for details.
Example I.
Example II.
Example III.
Input image

DPM output

Our output

3D models
human evaluation

Orientation quality at 25% recall

<table>
<thead>
<tr>
<th></th>
<th>Good</th>
<th>Bad</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exemplar-LDA</td>
<td>52%</td>
<td>48%</td>
</tr>
<tr>
<td>Ours</td>
<td>90%</td>
<td>10%</td>
</tr>
</tbody>
</table>
human evaluation

Style consistency at 25% recall

<table>
<thead>
<tr>
<th></th>
<th>Exact</th>
<th>Ok</th>
<th>Bad</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exemplar-LDA</td>
<td>3%</td>
<td>31%</td>
<td>66%</td>
</tr>
<tr>
<td>Ours</td>
<td>21%</td>
<td>64%</td>
<td>15%</td>
</tr>
</tbody>
</table>
3D Object Manipulation in a Single Photograph using Stock 3D Models

Natasha Kholgade\textsuperscript{1} \quad Tomas Simon\textsuperscript{1} \quad Alexei Efros\textsuperscript{2} \quad Yaser Sheikh\textsuperscript{1}

\textsuperscript{1}Carnegie Mellon University \quad \textsuperscript{2}University of California, Berkeley
3D Object Manipulation in a Single Photograph using Stock 3D Models

Natasha Kholgade¹  Tomas Simon¹  Alexei Efros²  Yaser Sheikh¹

¹Carnegie Mellon University  ²University of California, Berkeley
The Language Bottleneck

Mental Picture

words

Image