Teaching Computers to See Using Big 3D Data

Jianxiong Xiao

PRINCETON VISION GROUP
Game

I’ll show a picture for 0.1 second. Tell me what you see.
Scene Understanding
Scene Understanding

CSI: Crime Scene Investigation
Scene Understanding

1 cat, 2 people, 3 cars, and grass.
Scene Understanding
State-of-the-Art: Deep Learning

Scientists See Promise in Deep-Learning Programs

Deep Learning

With massive amounts of computational power, machines can now recognize objects and translate speech in real time. Artificial intelligence is finally getting smart.
NYU Algorithm on NYU Dataset

---

### OverFeat:
Integrated Recognition, Localization and Detection using Convolutional Networks

Pierre Sermanet, David Eigen, Xiang Zhang, Michael Mathieu, Rob Fergus, Yann LeCun
Courant Institute of Mathematical Sciences, New York University
719 Broadway, 12th Floor, New York, NY 10003
sermanet,deigen,xiang.mathieu.fergus,yann@cs.nyu.edu

---

### Indoor Segmentation and Support Inference from RGBD Images

Nathan Silberman\(^1\), Derek Hoiem\(^2\), Pushmeet Kohli\(^3\), Rob Fergus\(^4\)

\(^1\)Courant Institute, New York University
\(^2\)Department of Computer Science, University of Illinois at Urbana-Champaign
\(^3\)Microsoft Research, Cambridge

---

Rob Fergus
NYU

Yann LeCun
NYU
NYU algorithm on NYU dataset

5 most likely categories:
0.236223 shoe shop, shoe-shop, shoe store
0.027985 confectionery, confectionary
0.025233 cinema, movie theater
0.024637 butcher shop, meat market
0.024317 slot, one-armed bandit

NYU1418.jpg
NYU algorithm on NYU dataset

5 most likely categories:
- 0.181371 potter's wheel
- 0.175774 chiffonier, commode
- 0.126224 stove
- 0.060575 pedestal, plinth, footstall
- 0.044722 file, file cabinet, filing cabinet
NYU algorithm on NYU dataset

Most likely categories:
- 0.094286 schipperke
- 0.086387 Labrador retriever
- 0.057771 black-and-tan coonhound
- 0.0531486 Staffordshire bullterrier
- 0.033145 curly-coated retriever
Why is Vision So Hard?
Structure of Ambient Light

Animation from A. Torralba
Structure of Ambient Light
Viewpoint

Image

3D world

3D world

3D world
But
3D Sensors

- Microsoft Kinect
- Intel RealSense
- Google Project Tango
- Apple Primesense
- Asus Xtion
- LEAP Motion
- Structure.io
- Stereo Cameras
Depth for Other Vision Tasks

Kinect Fusion

Newcombe et al. 2011

Intrinsic Image

Barron & Malik, 2013

Human Pose Recognition

Shotton et al. 2011
Sliding Shapes

Input: Kinect Depth Map

Output: 3D Bounding Box

S. Song, J. Xiao
Sliding Shapes for 3D Object Detection in Depth Images
ECCV 2014 Oral
Depth-based Object Detection

[0.765] Sliding Shapes
Depth-based Object Detection

\[ \text{chair} \]

\[ \times 1.7 \text{ improvement on Average Precision compared to the best of DPM & R-CNN} \]
Algorithm
Training: CAD models
Training: CAD models
Training: Rendering Depth
Training: 3D Exemplar SVM

- Rendered Depth
- Point Cloud
- Feature

Linear SVM Classifier
3D Features

Points  Normal  Shape  TSDF  Combined
Testing: 3D Sliding Window
Testing: 3D Sliding Window
Testing: 3D Sliding Window

SVM1

No

SVM2

No

SVM3

No
Testing: 3D Sliding Window

SVM1
No

SVM2
No

SVM3
No
Testing: 3D Sliding Window

SVM1

SVM2

SVM3

No

No

No
Testing: 3D Sliding Window

SVM1: No
SVM2: Yes
SVM3: No
Testing: 3D Sliding Window

SVM1: Yes
SVM2: Yes
SVM3: No
Testing: 3D Sliding Window
Results
Results
Results
## Evaluation

### Sliding Shapes

<table>
<thead>
<tr>
<th></th>
<th>chair</th>
<th>toilet</th>
<th>bed</th>
<th>sofa</th>
<th>table</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D+</td>
<td>0.316</td>
<td>0.331</td>
<td>0.381</td>
<td>0.315</td>
<td>0.289</td>
</tr>
<tr>
<td>3D</td>
<td>0.765</td>
<td>0.749</td>
<td>0.741</td>
<td>0.403</td>
<td>0.474</td>
</tr>
<tr>
<td>2D+</td>
<td>0.643</td>
<td>0.644</td>
<td>0.412</td>
<td>0.339</td>
<td>0.314</td>
</tr>
<tr>
<td>2D</td>
<td>0.736</td>
<td>0.736</td>
<td>0.751</td>
<td>0.418</td>
<td>0.478</td>
</tr>
</tbody>
</table>

### DPM best

<table>
<thead>
<tr>
<th></th>
<th>chair</th>
<th>toilet</th>
<th>bed</th>
<th>sofa</th>
<th>table</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D+</td>
<td>0.176</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3D</td>
<td>0.446</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2D+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2D</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### RCNN VOC

<table>
<thead>
<tr>
<th></th>
<th>chair</th>
<th>toilet</th>
<th>bed</th>
<th>sofa</th>
<th>table</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.200</td>
</tr>
<tr>
<td>3D</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.203</td>
</tr>
<tr>
<td>2D+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2D</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

**RGB**

<table>
<thead>
<tr>
<th></th>
<th>chair</th>
<th>toilet</th>
<th>bed</th>
<th>sofa</th>
<th>table</th>
</tr>
</thead>
<tbody>
<tr>
<td>DPM-VOC[2]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.213</td>
</tr>
<tr>
<td>DPM-SUN[2]</td>
<td>-</td>
<td>0.131</td>
<td>0.345</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DPM-RMRC[2]</td>
<td>-</td>
<td>0.115</td>
<td>0.269</td>
<td>0.344</td>
<td>0.318</td>
</tr>
<tr>
<td>RCNN-VOC[6]</td>
<td>-</td>
<td>0.182</td>
<td>0.342</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Comparison with DPM

Sliding Shapes

DPM
Comparison with DPM

Sliding Shapes

DPM
Comparison with DPM

Sliding Shapes

DPM
Comparison with DPM

Sliding Shapes

DPM
Comparison with DPM

Sliding Shapes

DPM
Analysis
Object Detection is Hard

"These techniques are inadequate for three-dimensional scene analysis for many reasons:"

1. Variation
2. Viewpoint
3. Illumination
4. Clutter
5. Occlusion

Jia Deng’s advisor’s advisor’s advisor’s advisor


Imagenet
How does Big Data Helps?
1. Variation

2. Viewpoint

3. Illumination

4. Clutter

5. Occlusion
Problem: Intra-class Variations

1. Variation
2. Viewpoint
3. Illumination
4. Clutter
5. Occlusion

CAD Models Used for Training
Solution: Data-driven Exemplar

Solution: Data-driven Exemplar

1. Variation
2. Viewpoint
3. Illumination
4. Clutter
5. Occlusion

AP

Number of Models

- chair
- bed
- toilet
- sofa
- table

0 0.2 0.4 0.6 0.8 1

0 1 2 3 4 5 6 7 8 9 10 11
Problem: Viewpoints

1. Variation
2. Viewpoint
3. Illumination
4. Clutter
5. Occlusion
Solution: Numerate All Views

1. Variation

2. Viewpoint

3. Illumination

4. Clutter

5. Occlusion
1. Variation

2. Viewpoint

3. Illumination

4. Clutter

5. Occlusion
Solution: Numerate All Views

1. Variation
2. Viewpoint
3. Illumination
4. Clutter
5. Occlusion
How does 3D Data Helps?
Problem: Illumination

- Color Rendering ≠ Real Photo
Solution: 3D Depth

- Color Rendering $\neq$ Real Photo
- Depth Rendering $\approx$ Depth from Kinect

Kinect Body Pose Recognition [Shotton et al.]
Problem: Clutter

3D mesh
Solution: Occupation Mask

3D mesh  →  Occupation mask
Solution: Occupation Mask

Clutter

Don’t Care

Feature

SVM

1. Variation

2. Viewpoint

3. Illumination

4. Clutter

5. Occlusion
Solution: Occupation Mask

Comparison data comes here. How important is the occupation mask?

Without:

0.689

With:

0.765

62

0
0.2
0.4
0.6
0.8
1

0
0.2
0.4
0.6
0.8
1

Precision

Recall

[0.765] with

[0.698] without

1. Variation

2. Viewpoint

3. Illumination

4. Clutter

5. Occlusion
Problem: Occlusion
Solution: 3D Window

Using depth, we can know which part is **occluded**. In 3D, we can separate the object from the **occluder**.
False Positives
False Positives
False Positives
Misses
Big 3D Data

1. For Bottom-up Detection
Big 3D Data

1. For Bottom-up Detection

2. For Top-down Context
The Context Challenge
Improvement on PASCAL <1.5%
A Typical Bedroom
What Your Eyes See
What a Camera Sees

focal length = 35 mm
Unfair?
What is this object?
What is this object?
What is this object?
What is this object?
What is this object?
What is this object?
What is this object?
What is this object?

Look-Alikes by Joan Steiner
Small Field-of-view

1. Small number of objects $\rightarrow$ little interplay.

   1.5 object classes
   2.7 object instances

2. The occurrence of objects is unpredictable.
PanoContext

input

output

PanoContext: A Whole-room 3D Context Model for Panoramic Scene Understanding
Y. Zhang, S. Song, P. Tan and J. Xiao
ECCV 2014 Oral
Input: A Panorama
Output: 3D Scene Parsing
Output: 3D Scene Reconstruction
2-Step Algorithm

1. Generation a pool of hypotheses
2. Choose the best hypothesis

2-Step Algorithm

\[ f(\cdot) \]

Hypotheses

\[ \checkmark \]

\[ \times \]

\[ \cdots \]

\[ \times \]
Context in 3D
Is it a valid room?

bedroom
Is it a valid room?

Pairwise? Hierarchical? ....?

Gaussian? Dirichlet? ....?
Is it a valid room?
The ultimate solution for all problems in the world:

Nearest Neighbor
Is it a valid room?

<table>
<thead>
<tr>
<th>Training Data 1</th>
<th>Training Data 2</th>
<th>Training Data 3</th>
<th>Training Data 4</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image3.png" alt="Image" /></td>
<td><img src="image4.png" alt="Image" /></td>
</tr>
<tr>
<td>difference=</td>
<td>difference=</td>
<td>difference=</td>
<td>difference=</td>
</tr>
<tr>
<td>0.91</td>
<td>0.20</td>
<td>0.45</td>
<td>1.32</td>
</tr>
</tbody>
</table>
3D Annotated Panorama Dataset

539 bedrooms  448 living rooms  317 bathrooms

http://panocontext.cs.princeton.edu
Transform GT $\rightarrow$ Big 3D Data

Ground truth

Resize Y
0.84
0.83
0.35
0.25
0.83

Resize X
0.86
0.70
0.20
0.22
1.21

Fix dist. to wall
0.73
0.52
0.33
0.52
0.97

Rotation & scale
1.42
2.82
4.23
4.44
1.62
Is it a valid room?

<table>
<thead>
<tr>
<th>Training Data 1</th>
<th>Training Data 2</th>
<th>Training Data 3</th>
<th>Training Data 4</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image3.png" alt="Image" /></td>
<td><img src="image4.png" alt="Image" /></td>
</tr>
<tr>
<td><strong>Hypothesis</strong></td>
<td><strong>Hypothesis</strong></td>
<td><strong>Hypothesis</strong></td>
<td><strong>Hypothesis</strong></td>
</tr>
<tr>
<td><img src="graph1.png" alt="Graph" /></td>
<td><img src="graph2.png" alt="Graph" /></td>
<td><img src="graph3.png" alt="Graph" /></td>
<td><img src="graph4.png" alt="Graph" /></td>
</tr>
<tr>
<td>0.91</td>
<td>0.20</td>
<td>0.45</td>
<td>1.32</td>
</tr>
</tbody>
</table>
Bedroom
Bedroom
Living Room
Living Room
Data-driven 3D Context Helps?
3D Context vs. 2D Appearance

- Context is as powerful as local appearance for object detection.
3D Context vs. 2D Appearance

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

![Graph showing precision-recall curves for different scenarios: bed, painting, desk, tv, chair. The curves for DPM, PanoContext, and Context+Detector are compared.](image)
How does data-driven 3D context help?

- Helps to decide 3D sizes of objects

DPM: Wrong relative size

PanoContext
How does data-driven 3D context help?

- Helps to decide 3D sizes of objects
- Helps to decide number of objects

DPM: Wrong number of objects

Our detection
How does data-driven 3D context help?

- Helps to decide 3D sizes of objects
- Helps to decide number of objects
- Helps to constrain relative position

DPM: Wrong relative position

Our detection
Big 3D Data

1. For Bottom-up Detection

2. For Top-down Context
Big 3D Data

1. For Bottom-up Detection

2. For Top-down Context

3. For Feature/Shape Representation
3D Feature/Representation

Points  Normal  Shape  TSDF  Combined
3D Shape Representation
Life is just a matter of perspective!
3D Shape Representation
3D Shape Representation?

Theory

Geon

Generalized Cylinder

The state-of-the-arts

Dalal & Triggs 2005

Felzenszwalb et al. 2010

Instance-level Matching

Rothganger et al. 2006

Philbin et al. 2007

Girshick et al. 2014

R-CNN: Regions with CNN features

1. Input image
2. Extract region proposals (~2k)
3. Compute CNN features
4. Classify regions
Data-driven 3D Feature/3D Shape Representation

• Learn from Big 3D Data (not hand-defined)
• Able to generalize (not just memorize NN)
• Compositional (from parts to whole object)
• Beyond recognition (e.g. shape completion, NBV)

Z. Wu, S. Song, A. Khosla, X. Tang, J. Xiao
3D ShapeNets for 2.5D Object Recognition and Next-Best-View Prediction
arXiv:1406.5670 [cs.CV]
3D ShapeNets

Convolu7onal Deep Belief Network $p(x, y)$

Visualization at Different Layers
As a 3D Shape Prior

Sampled Models
### As a 3D Feature Extractor

<table>
<thead>
<tr>
<th>Feature</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>spherical harmonic</td>
<td>82.0%</td>
</tr>
<tr>
<td>light field</td>
<td>86.1%</td>
</tr>
<tr>
<td>ours: 5th layer</td>
<td>86.5%</td>
</tr>
<tr>
<td>ours: 6th layer</td>
<td>83.7%</td>
</tr>
<tr>
<td>ours: 7th layer</td>
<td>82.0%</td>
</tr>
</tbody>
</table>

**Mesh Classification**

**Mesh Retrieval**

---

[Graph showing precision-recall curves for different features and layers.]
Shape Completion

Depth  Truth  3D ShapeNets  NN
2.5D Object Recognition

\[ \mathbf{x} = (\mathbf{x}_u, \mathbf{x}_o) \quad p(y|x_o) \quad \text{Gibbs sampling with clamping} \]
2.5D Object Recognition

Depth map from the back of a sofa

Volumetric representation

What is it?

sofa?
dresser?
bathtub?

Not sure. Look from another view?

Where to look next?

Next-Best-View

3D ShapeNets

Aha! It is a sofa!

New depth map
Big 3D Data

1. For Bottom-up Detection
   - Sliding Shapes

2. For Top-down Context
   - PanoContext

3. For Feature/Shape Representation
   - 3D ShapeNets
3D is the “near” Future

• Niloy:
  Present a vision as to where the field is heading

• My two cents:
  Learning 3D representation from 3D data
  – Not much left from 2D image feature learning (CNN)
  – A lot to be done for 3D feature, context, reasoning
  – 3D data is finally here!
    • RGB-D sensors
    • Online 3D repositories
Big 3D Data is a Beast

Big Data + 3D Data = Big 3D Data
How to master Big 3D Data?
Key Challenges to Tackle

- A set of algorithms on different paradigms
  - 3D Deep Learning
  - 3D Data-driven Brute-force NN
  - 3D Probabilistic Grammar for Context

3D ShapeNets  PanoContext  Liu et al. 2014
Key Challenges to Tackle

• A set of algorithms on different paradigms
  – 3D Deep Learning
  – 3D Data-driven Brute-force NN
  – 3D Probabilistic Grammar for Context

• Being More Focus and Scientific
  – Standard dataset
  – Standard tasks
Princeton ModelNet
Princeton ModelNet

585 categories 127,915 CAD models  http://modelnet.cs.princeton.edu
ShapeNet

ShapeNet 2014 Fall Preview
ShapeNet 2014 Fall Preview

Taxonomy
- ShapeNet 2014 Fall Preview
  - natural object(7,1006)
  - celestial body, heav
  - covering, natural cov
  - extraterrestrial obje
  - mechanism(1, 1)
  - radiator(0, 116)
  - rock, stone(0, 59)
  - plant part, plant stru
  - artifact, artefact(26, 1206)
  - sport, athletics(14, 1764)
  - geological formation, for
  - Misc(319, 17175)

Synset models
- Displaying 1 to 40 of 134581

Pat Hanrahan
Leonidas Guibas
Thomas Funkhouser
Silvio Savarese
SUN3D Database

http://sun3d.cs.princeton.edu
RGB-D Tracking Benchmark

Tracking Revisited using RGBD Camera: Unified Benchmark and Baselines.
S. Song and J. Xiao
ICCV2013.
✓ PASCAL-scale size

✓ All densely labeled
  • 2D segmentation
  • 3D object label
  • 3D object orientation
  • 3D room layout
  • scene type

SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite.
S. Song, S. Lichtenberg, L. Zhang, F. Yu, Y. Zhang and J. Xiao
Under Review.
SUN RGB-D

Kinect v2 mounted on stabilizer

Battery and cables in backpack

Mobile laptop harness holds the computer

10,000 RGB-D Scene Images

Intel Realsense

Asus Xtion

Kinect v1

Kinect v2
SUN RGB-D

2D segmentation

3D annotation

bedroom

classroom

conference room
SUN RGB-D
SUN RGB-D
SUN RGB-D Benchmark Suite

- **color image**
- **depth image**

- **Scene Categorization**
- **Semantic Segmentation**
- **Room Layout Estimation**

- **home office**
  - lamp
  - wall
  - desk
  - chair
  - floor

- **Wall**
- **Floor**
SUN RGB-D Benchmark Suite

3D Object Detection
SUN RGB-D Benchmark Suite

table  
chair  
bookshelf

3D Object Detection
SUN RGB-D Benchmark Suite

3D Object Orientation

table
chair
bookshelf
Big 3D Data is a Way of Thinking

1. For Bottom-up Detection

2. For Top-down Context

3. For Feature/Shape Representation

4. Benchmark Suite

SUN RGB-D