Teaching Computers to See Using Big 3D Data

Jianxiong Xiao
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

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.

The New York Times

Scientists See Promise in Deep-Learning Programs
NYU Algorithm on NYU Dataset

OverFeat:
Integrated Recognition, Localization and Detection using Convolutional Networks

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

Indoor Segmentation and Support Inference from RGBD Images

Nathan Silberman¹, Derek Hoiem², Pushmeet Kohli³, Rob Fergus¹

¹Courant Institute, New York University
²Department of Computer Science, University of Illinois at Urbana-Champaign
³Microsoft 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
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:
- 0.09286 schipperke
- 0.08887 Labrador retriever
- 0.05771 black-and-tan coonhound
- 0.053486 Staffordshire bullterrier
- 0.033145 curly-coated retriever

NYU0523.jpg
Why is Vision So Hard?
Structure of Ambient Light
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

×1.7 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

SVM1: No
SVM2: No
SVM3: No
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: No
SVM2: No
SVM3: 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

Exemplar Rendering
Results
Results
Results
Evaluation

- Sliding Shapes
- DPM best
- RCNN VOC

<table>
<thead>
<tr>
<th>Method</th>
<th>chair</th>
<th>toilet</th>
<th>bed</th>
<th>sofa</th>
<th>table</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sliding Shapes</td>
<td>0.316</td>
<td>0.765</td>
<td>0.331</td>
<td>0.749</td>
<td>0.643</td>
</tr>
<tr>
<td>3D+</td>
<td>0.743</td>
<td>0.643</td>
<td>0.736</td>
<td>0.644</td>
<td>0.736</td>
</tr>
<tr>
<td>3D</td>
<td>0.381</td>
<td>0.741</td>
<td>0.412</td>
<td>0.751</td>
<td>0.315</td>
</tr>
<tr>
<td>2D</td>
<td>0.163</td>
<td>0.213</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RGB-DPM-VOC[2]</td>
<td>0.176</td>
<td>0.446</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RGB-DPM-SUN[2]</td>
<td>0.131</td>
<td>0.345</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RGB-DPM-RMRC[2]</td>
<td>0.115</td>
<td>0.269</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RGB-RCNN-VOC[6]</td>
<td>0.182</td>
<td>0.342</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RGB-RCNN-VOC[6]</td>
<td>0.182</td>
<td>0.342</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Precision-Recall curves for different methods and objects (chair, toilet, bed, sofa, table) are shown.
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
Analysis
Object Detection is Hard

“The techniques are inadequate for three-dimensional scene analysis for many reasons:”

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

How does Big Data Helps?
1. Variation

2. Viewpoint

3. Illumination

4. Clutter

5. Occlusion
Problem: Intra-class Variations

CAD Models Used for Training
Solution: Data-driven Exemplar

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

![Graph showing the AP (Average Precision) for different numbers of models for various objects like chair, bed, toilet, sofa, and table.](graph.png)
Problem: Viewpoints
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

Feature

Don’t Care

SVM
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
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?

Spider
Plugin
Walkman
Gum
Nuts
CoQon	
  bud
Lid	
  of	
  can	
  food
Caps
Cotton bud
Lid of can food
Walkman

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

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

2. The occurrence of objects is unpredictable.

1.5 object classes
2.7 object instances
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

✓

✗

✗

✗
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="training_data_1.png" alt="Image" /></td>
<td><img src="training_data_2.png" alt="Image" /></td>
<td><img src="training_data_3.png" alt="Image" /></td>
<td><img src="training_data_4.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
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 of Training Data 1" /></td>
<td><img src="image2.png" alt="Image of Training Data 2" /></td>
<td><img src="image3.png" alt="Image of Training Data 3" /></td>
<td><img src="image4.png" alt="Image of Training Data 4" /></td>
</tr>
<tr>
<td><img src="graph1.png" alt="Graph of Hypothesis" /></td>
<td><img src="graph2.png" alt="Graph of Hypothesis" /></td>
<td><img src="graph3.png" alt="Graph of Hypothesis" /></td>
<td><img src="graph4.png" alt="Graph of Hypothesis" /></td>
</tr>
<tr>
<td>0.91</td>
<td>0.20</td>
<td>0.45</td>
<td>1.32</td>
</tr>
</tbody>
</table>

Hypothesis

Training Data 1

Training Data 2

Training Data 3

Training Data 4
Bedroom
Bedroom
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 methods: DPM, PanoContext, and Context+Detector. The graph includes a plot for 'bed' with precision on the y-axis and recall on the x-axis. Additional plots for 'painting', 'desk', 'tv', and 'chair' are also shown.](image)
How does data-driven 3D context help?

• Helps to decide 3D sizes of objects
How does data-driven 3D context help?

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

Instance-level Matching

Rothganger et al. 2006

Philbin et al. 2007

Felzenszwalb et al. 2010

The state-of-the-arts

Dalal & Triggs 2005

R-CNN: Regions with CNN features

1. Input image

2. Extract region proposals (~2k)

3. Compute CNN features

4. Classify regions

Rothganger et al. 2006

Philbin et al. 2007

Girshick et al. 2014

The set of geons is generated by variations in the production function for generalized cylinders that produce viewpoint-invariant (=nonaccidental) shape differences:

1. Cross Section: Straight vs. Curved
2. Axis: Straight vs. Curved
3. Size of Cross Section: Constant (parallel sides) vs. Expand vs. Expand & Contract vs Contract & Expand
4. Termination of Geon when Nonparallel: Truncated vs. Painted vs. Rounded

Geon Theory

Instance-level Matching

R-CNN: Regions with CNN features

Girshick et al. 2014
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

To study shape representation and view planning for recognition, we propose to represent a geometric 3D shape as a probability distribution of binary variables on a 3D voxel grid. Each 3D mesh is represented as a binary tensor: 1s indicate the voxels on or inside the mesh surface, and 0s indicate the voxels outside the mesh (i.e. empty space). The grid size in our experiments is $48 \times 48 \times 48$.

To represent the probability distribution of these binary variables for 3D shapes, we designed a Convolutional Deep Belief Network (CDBN). Deep Belief Network (DBN) [9] is a powerful probabilistic model for binary variables that is typically used for 2D images where they modeled the joint probabilistic distribution over pixels and labels. However, adapting the model from 2D pixel data to 3D voxel data is non-trivial. A 3D voxel volume with reasonable resolution (say $48 \times 48 \times 48$) would have the same dimension of a high-res image ($332 \times 332$). A fully connected DBN would result in a huge number of parameters that are intractable to train effectively. Therefore, we propose to use convolution to reduce model parameters by weight sharing. But different from typical convolutional deep learning model (e.g. [14]), we don't introduce any kind of pooling in hidden layers, because although pooling may give us invariance properties for recognition, it will also give us more uncertainty during reconstruction, which is important for next-best-view prediction.

The energy of a convolutional layer in our model is defined as:

$$E(v, h) = \sum_f \sum_j \langle h_f j \rangle W_f \times v_j + c_f h_f \langle \sum_l b_l v_l \rangle$$  \hspace{1cm} (1)$$

where $v_l$ denotes each visible unit, $h_f j$ denotes each hidden unit in a feature channel $f$, $W_f$ denotes the convolutional filter. The "\times" sign represents convolution operation. In this energy definition, each visible unit $v_l$ is associated with a unique bias term $b_l$ to facilitate reconstruction, and all hidden units $\{h_f j\}$ in the same convolution channel share the same bias term $c_f$. We also allow convolution stride as in [12].

A 3D shape is represented as a $42 \times 42 \times 42$ voxel grid with 3 extra cells of empty space on both direction for padding to reduce convolution border artifacts. The labels are presented as standard one of $K$ softmax variables. All of our training data are manually aligned to the same direction. The final architecture of our model is in Figure 2(a). The first layer has 80 filters of size 6 and stride 3;
Princeton ModelNet
Princeton ModelNet

585 categories
127,915 CAD models
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
Shape Completion

Depth  Truth  3D ShapeNets  NN
2.5D Object Recognition

Depth map from the back of a sofa

Volumetric representation

What is it?

sofa?
dresser?
bathtub?

3D ShapeNets

Not sure. Look from another view?

Next-Best-View

Where to look next?

New depth map

Aha! It is a sofa!
Big 3D Data

1. For Bottom-up Detection

2. For Top-down Context

3. For Feature/Shape Representation
Big 3D Data is a Beast

Big Data + 3D Data = Big 3D Data
IT staffs
How to master Big 3D Data?
Challenges for Big 3D Data

- Visualization
- Labeling
- Registration
- Benchmarking
- Dataset Construction
- ...

126
Navigate in Big 3D Data

Frick Collection Gallery (New York City)
Visualization

5x Google Streetview

1x Reconstructing the World's Museums, ECCV2012.
Labeling

SUN3D: A Database of Big Spaces Reconstructed using SfM and Object Labels. ICCV2013.
Registration

Bad Reconstruction

errors

Good Reconstruction

Semantic Object Labeling

Semantic Segmentation

SUN3D: A Database of Big Spaces Reconstructed using SfM and Object Labels. ICCV2013.
SUN3D Database

http://sun3d.cs.princeton.edu
Benchmarking

Tracking Revisited using RGBD Camera: Unified Benchmark and Baselines.
S. Song and J. Xiao
ICCV2013.
Big RGB-D Dataset

Shuran Song  Samuel Lichtenberg

2D Polygon Annotation  3D Bounding Box Annotation
Big 3D Data is a Way of Thinking

1. For Bottom-up Detection
   - Sliding Shapes

2. For Top-down Context
   - PanoContext

3. For Feature/Shape Representation
   - 3D ShapeNets

4. Challenges