The Promise and Peril of
Big Data

Alexei (Alyosha) Efros
CMU (school-year), INRIA (summer)
Outline

• Why we need lots of data?
• The Promise of Big Data
• Perils of Big Data
  – Bias
  – Long tails -- we will never have enough data. “Unfamiliar is common”
  – Categorization in the modern world: "Everything is Miscellaneous"
Texture

- Texture depicts spatially repeating patterns
- Many natural phenomena are textures

radishes  rocks  yogurt
Texture Synthesis
Classical Texture Synthesis

Synthesis

Parametric Texture Model

Analysis

Sample texture

Novel texture

This is hard!
Motivation from Language

• [Shannon, ’48] proposed a way to generate English-looking text using N-grams:
  – Assume a generalized Markov model
  – Use a large text to compute prob. distributions of each letter given N-1 previous letters
  – Starting from a seed repeatedly sample this Markov chain to generate new letters
  – Also works for whole words

WE NEED TO EAT CAKE
Mark V. Shaney (Bell Labs)

• Results (using alt.singles corpus):
  – “As I've commented before, really relating to someone involves standing next to impossible.”
  – “One morning I shot an elephant in my arms and kissed him.”
  – “I spent an interesting evening recently with a grain of salt”

• Notice how well local structure is preserved!
  – Now, instead of letters let’s try pixels…”
Non-parametric Approach

Synthesis

Novel texture

Analysis

Sample texture
[Efros & Leung, ’99]

non-parametric sampling

Input image
Texture Growing
Two Kinds of Things in the World

Navier-Stokes Equation

\[ \frac{\partial \mathbf{u}}{\partial t} = -(\mathbf{u} \cdot \nabla) \mathbf{u} + \nu \nabla^2 \mathbf{u} - \frac{1}{\rho} \nabla p + \mathbf{f} \]

+ weather
+ location
+ ...
Lots of data available
“Unreasonable Effectiveness of Data”  
[Halevy, Norvig, Pereira 2009]

• Parts of our world can be explained by elegant mathematics:
  – physics, chemistry, astronomy, etc.

• But much cannot:
  – psychology, genetics, economics, etc.

• Enter: The Magic of **Big Data**
  – Great advances in several fields:
    • e.g. speech recognition, machine translation, Google
• A.I. for the postmodern world:
  – all questions have already been answered…many times, in many ways
  – Google is dumb, the “intelligence” is in the data
Computer Vision

Two disciplines which happen to share the same name:

- **Vision as Measurement:**
  - e.g. stereo, structure-from-motion, illumination estimation
  - output: depth (meters), visual angle (radians), brightness (cd/m^2), etc.

- **Vision as Understanding:**
  - recognition
  - output: human concepts
Recognition Learning Spectrum

Extrapolation problem
Generalization

Interpolation problem
Correspondence

Number of training samples

Traditional datasets

Slide by Antonio Torralba
Face Detection: Big Success Story

- Rowley, Baluja, and Kanade, 1998
- Schniderman & Kanade, 1999
- Viola & Jones, 2001
Modern Recognition is largely Data-Driven

• In non-linear SVMs:
  – In ML, people report ~10% of data are support vectors
  – In recognition, up to 2/3 of data are support vectors!!!

• In linear SVMs:
  – Typical setup: 2000 dim. HOG, only 300 “chair” examples

Figure that Francis Bach hates
Everything else being equal...

... the visual world is just much richer!

- **MNIST Digits**
  - 10 digits *
  - ~1,000 variations = 10,000

- **English words**
  - ~100,000 words *
  - ~5 variations = 500,000

- **Visual world**
  - ~100,000 objects *
  - ~10,000 variations (pose, scale, lighting, intra-category)
  - = 1,000,000,000 (1 billion!)
Yet, we train on 15 examples?!
my claim:

Large-scale data is necessary, but certainly not sufficient, to solve recognition
80 Million Tiny Images

Visual dictionary
Click on top of the map to visualize the images in that region of the visual dictionary.

Torralba, Fergus, Freeman. PAMI 2008
How connected is the visual space?

Torralba, Fergus, Freeman. PAMI 2008
Magic Of data

7,900

Target

A. Torralba, R. Fergus, W.T.Freeman. PAMI 2008
Magic Of data

A. Torralba, R. Fergus, W.T. Freeman. PAMI 2008
Magic
Of
data
Is this something humans do at all?
What’s the Capacity of Visual Long Term Memory?

What we know...

Standing (1973)
10,000 images
83% Recognition

... people can remember thousands of images

What we don’t know...

... what people are remembering for each item?

According to Standing

“Basically, my recollection is that we just separated the pictures into distinct thematic categories: e.g. cars, animals, single-person, 2-people, plants, etc.) Only a few slides were selected which fell into each category, and they were visually distinct.”

Slide by Aude Oliva
Showed 14 observers 2500 categorically unique objects
1 at a time, 3 seconds each
800 ms blank between items
Study session lasted about 5.5 hours
Repeat Detection task to maintain focus
Followed by 300 2-alternative forced choice tests
how far can we push the fidelity of visual LTM representation?

Same object, different states

Slide by Aude Oliva
Massive Memory I: Recognition Memory Results

Replication of Standing (1973)

92%
Massive Memory I: Recognition Memory Results

- **Novel**: 92%
- **Exemplar**: 88%
- **State**: 87%

Slide by Aude Oliva
The Good News

Really stupid algorithms + Lots of Data

= “Unreasonable Effectiveness”
Raw (unlabelled) Data

Useful since visual world has structure

# actual images << # possible images

Number of images seen by all humanity: \(10^{20}\)

106,456,367,669 humans \(\times\) 60 years \(\times\) 3 images/second \(\times\) 60 \(\times\) 60 \(\times\) 16 \(\times\) 365 =


Number of all 32x32 images:

\(256^{32 \times 32} \approx 10^{7373}\)
Automatic Colorization Result

Grayscale input High resolution

Colorization of input using average

A. Torralba, R. Fergus, W.T.Freeman. 2008
Automatic Orientation

• Many images have ambiguous orientation

• Look at top 25% by confidence:

• Examples of high and low confidence images:

Slide by Antonio Torralba
Image Restoration using Online Photo Collections [ICCV’09]
[Hays & Efros, SIGGRAPH’07]
Scene Descriptor
Scene Descriptor

Gist scene descriptor
(Oliva and Torralba 2001)
Scene Descriptor

Gist scene descriptor
(Oliva and Torralba 2001)
2 Million Flickr Images
... 200 scene matches
Nearest neighbors from a collection of 20 thousand images
Nearest neighbors from a collection of 2 million images
Scene matching with camera transformations

Sivic, Kaneva, Torralba, Avidan, Freeman, Internet Vision Workshop, 2008
Scene matching with camera view transformations: Translation

1. Move camera
2. View from the virtual camera
3. Find a match to fill the missing pixels
4. Locally align images
5. Find a seam
6. Blend in the gradient domain
Scene matching with camera view transformations: Camera rotation

1. Rotate camera

2. View from the virtual camera

3. Find a match to fill-in the missing pixels

4. Stitched rotation

5. Display on a cylinder
Josef’s cool movie...
Data with labels (correspondences)

Useful even for really noisy labels!
Two simple ways to use Lots of Data

- Find that needle in the haystack and disregard the rest (a.k.a. kNN)
- See what different subsets of data think of you
1. kNN + Label Transfer

Sky, Water, Hills, Beach, Sunny, mid-day
80 Million Tiny Images [PAMI’08]

Torralba, Fergus, Freeman, PAMI 2008
Non-parametric Scene Parsing

[CVPR’09]

Liu, Yuen, Torralba, CVPR 2009
im2gps [CVPR’08]

Query Photograph

Hays & Efros, CVPR 2008
Assembling Visual Content

• Semantic Photo Synthesis [Johnson, ’06]

• Photo Clip Art [Lalonde, ‘07]

• Sketch2Photo [Chen,’09]
2. Subpopulation Labels

e.g. See Attributes
Priors for Large Photo Collections
&
What they Reveal about Cameras

Sujit Kuthirummal  Columbia University
Aseem Agarwala  Adobe Systems, Inc.
Dan B Goldman  Adobe Systems, Inc.
Shree K. Nayar  Columbia University
Compute Aggregate Statistic

Independent of Scenes & Photographers

One Camera’s Distortion

Camera Distortion Free

Independent of Scenes, Photographers & Cameras

Recover Camera Properties

Compute Aggregate Statistic

Independent of Scenes & Photographers Dependent on Camera
3. Relative (e.g. binary) labels

- Many concepts lack precise definition
  - E.g. beauty
- Or well-established boundaries
  - E.g. are curtains furniture?
- Or definition is context-specific:
  - E.g. "hairy" in "hairy dog" vs. "hairy man"

- Relative attributes:
  - Same/different, degree of similarity, etc.
  - Work starting on this, e.g.:
will **Big Data** solve all your problems?
1. Data is Biased

• Internet is a tremendous repository of visual data (Flickr, YouTube, Picassa, etc)

• But it’s not random samples of visual world
Flickr Paris
My Paris
Real Notre Dame
Sampling Bias

- People like to take pictures on vacation
Photographer Bias

- People want their pictures to be recognizable and/or interesting

vs.

[Image of a building]
Social Bias

Mildred and Lisa

Source: U.S. Social Security Administration

Gallagher et al CVPR 2008
Social Bias

Gallagher et al, CVPR 2008

Gallagher et al, CVPR 2009
Brief History of Recognition Datasets

“trying to escape bias”
The first dataset
Columbia Object Image Library (COIL-100) (1996)
Corel Dataset
image + partial grouping ⇒ segmentation

Yu & Shi, 2004
Caltech 101

Description

Pictures of objects belonging to 101 categories. About 40 to 800 images per category. Most categories have about 50 images. Collected in September 2003 by Fei-Fei Li, Marco Andreetto, and Marc 'Aurelio Ranzato. The size of each image is roughly 300 x 200 pixels.
Visual Object Classes Challenge 2011 (VOC2011)
TinyImages + ImageNet

80 Million Tiny Images

Visual dictionary
Click on top of the map to visualize the images in that region of the visual dictionary.
Unbiased Look at Dataset Bias
Torralba & Efros, CVPR 2011

• How much does this bias affect standard datasets used for object recognition?
“Name That Dataset!” game

1. Caltech 101
2. Caltech 256
3. MSRC
4. UIUC cars
5. Tiny Images
6. Corel
7. PASCAL 2007
8. LabelMe
9. COIL-100
10. ImageNet
11. 15 Scenes
12. SUN’09
SVM plays “\textit{Name that dataset!}”

- 12 1-vs-all classifiers
- Standard full-image features
- 39% performance (chance is 8%)

<table>
<thead>
<tr>
<th></th>
<th>UIUC</th>
<th>LabelMe Spain</th>
<th>PASCAL 2007</th>
<th>MSRC</th>
<th>SUN09</th>
<th>15 Scenes</th>
<th>Corel</th>
<th>Caltech101</th>
<th>Caltech256</th>
<th>Tiny</th>
<th>ImageNet</th>
<th>COIL-100</th>
</tr>
</thead>
<tbody>
<tr>
<td>UIUC</td>
<td>0</td>
<td>29</td>
<td>8</td>
<td>21</td>
<td>10</td>
<td>2</td>
<td>17</td>
<td>6</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>LabelMe Spain</td>
<td>54</td>
<td>0</td>
<td>7</td>
<td>8</td>
<td>6</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PASCAL 2007</td>
<td>10</td>
<td>29</td>
<td>10</td>
<td>10</td>
<td>7</td>
<td>4</td>
<td>7</td>
<td>7</td>
<td>11</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>MSRC</td>
<td>0</td>
<td>3</td>
<td>7</td>
<td>4</td>
<td>1</td>
<td>8</td>
<td>7</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SUN09</td>
<td>14</td>
<td>9</td>
<td>9</td>
<td>24</td>
<td>17</td>
<td>11</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>15 Scenes</td>
<td>6</td>
<td>3</td>
<td>13</td>
<td>51</td>
<td>11</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Corel</td>
<td>1</td>
<td>2</td>
<td>6</td>
<td>8</td>
<td>11</td>
<td>35</td>
<td>10</td>
<td>7</td>
<td>7</td>
<td>9</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Caltech101</td>
<td>2</td>
<td>9</td>
<td>9</td>
<td>2</td>
<td>4</td>
<td>7</td>
<td>38</td>
<td>14</td>
<td>7</td>
<td>6</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Caltech256</td>
<td>2</td>
<td>8</td>
<td>8</td>
<td>10</td>
<td>18</td>
<td>20</td>
<td>11</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Tiny</td>
<td>1</td>
<td>2</td>
<td>8</td>
<td>6</td>
<td>4</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>24</td>
<td>12</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ImageNet</td>
<td>1</td>
<td>3</td>
<td>11</td>
<td>9</td>
<td>5</td>
<td>11</td>
<td>8</td>
<td>12</td>
<td>13</td>
<td>21</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>COIL-100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
SVM plays “Name that dataset!”

![Graph showing recognition performance vs. number of training examples per class for different features: Hog2x2, Gist, Color thumbnail, Gray thumbnail, and Chance. Each feature has a line with error bars indicating variability.]
Dataset look-alikes

ImageNet pretending to be:

Caltech 256 look-alikes from ImageNet
COREL look-alikes from ImageNet
MSRC look-alikes from ImageNet

PASCAL VOC pretending to be:

15 scenes look-alikes from PASCAL 2007
MSRC look-alikes from PASCAL 2007
Caltech 101 look-alikes from PASCAL 2007
Datasets have different goals…

• Some are object-centric (e.g. Caltech, ImageNet)

• Otherwise are scene-centric (e.g. LabelMe, SUN’09)

• What about playing “name that dataset” on bounding boxes?
Similar results

Performance: 61%
(chance: 20%)
Measuring Dataset Bias
Cross-Dataset Generalization

Classifier trained on MSRC cars
## Cross-dataset Performance

Table 1. Cross-dataset generalization. Object detection and classification performance (AP) for “car” and “person” when training on one dataset (rows) and testing on another (columns), i.e. each row is: training on one dataset and testing on all the others. “Self” refers to training and testing on the same dataset (same as diagonal), and “Mean Others” refers to averaging performance on all except self.

<table>
<thead>
<tr>
<th>task</th>
<th>Train on:</th>
<th>Test on:</th>
<th>SUN09</th>
<th>LabelMe</th>
<th>PASCAL</th>
<th>ImageNet</th>
<th>Caltech101</th>
<th>MSRC</th>
<th>Self</th>
<th>Mean others</th>
<th>Percent drop</th>
</tr>
</thead>
<tbody>
<tr>
<td>“car”</td>
<td>SUN09</td>
<td></td>
<td>28.2</td>
<td>29.5</td>
<td>16.3</td>
<td>14.6</td>
<td>16.9</td>
<td>21.9</td>
<td>28.2</td>
<td>19.8</td>
<td>30%</td>
</tr>
<tr>
<td></td>
<td>LabelMe</td>
<td></td>
<td>14.7</td>
<td>34.0</td>
<td>16.7</td>
<td>22.9</td>
<td>43.6</td>
<td>24.5</td>
<td>34.0</td>
<td>24.5</td>
<td>28%</td>
</tr>
<tr>
<td></td>
<td>PASCAL</td>
<td></td>
<td>10.1</td>
<td>25.5</td>
<td>35.2</td>
<td>43.9</td>
<td>44.2</td>
<td>39.4</td>
<td>35.2</td>
<td>32.6</td>
<td>7%</td>
</tr>
<tr>
<td></td>
<td>ImageNet</td>
<td></td>
<td>11.4</td>
<td>29.6</td>
<td>36.0</td>
<td>57.4</td>
<td>52.3</td>
<td>42.7</td>
<td>57.4</td>
<td>34.4</td>
<td>40%</td>
</tr>
<tr>
<td></td>
<td>Caltech101</td>
<td></td>
<td>7.5</td>
<td>31.1</td>
<td>19.5</td>
<td>33.1</td>
<td>96.9</td>
<td>42.1</td>
<td>96.9</td>
<td>26.7</td>
<td>73%</td>
</tr>
<tr>
<td></td>
<td>MSRC</td>
<td></td>
<td>9.3</td>
<td>27.0</td>
<td>24.9</td>
<td>32.6</td>
<td>40.3</td>
<td>68.4</td>
<td>68.4</td>
<td>26.8</td>
<td>61%</td>
</tr>
<tr>
<td>Mean others</td>
<td></td>
<td></td>
<td>10.6</td>
<td>28.5</td>
<td>22.7</td>
<td>29.4</td>
<td>39.4</td>
<td>34.1</td>
<td>53.4</td>
<td>27.5</td>
<td>48%</td>
</tr>
<tr>
<td>“car”</td>
<td>SUN09</td>
<td></td>
<td>69.8</td>
<td>50.7</td>
<td>42.2</td>
<td>42.6</td>
<td>54.7</td>
<td>69.4</td>
<td>69.8</td>
<td>51.9</td>
<td>26%</td>
</tr>
<tr>
<td></td>
<td>LabelMe</td>
<td></td>
<td>61.8</td>
<td>67.6</td>
<td>40.8</td>
<td>38.5</td>
<td>53.4</td>
<td>67.0</td>
<td>67.6</td>
<td>52.3</td>
<td>23%</td>
</tr>
<tr>
<td></td>
<td>PASCAL</td>
<td></td>
<td>55.8</td>
<td>55.2</td>
<td>62.1</td>
<td>56.8</td>
<td>54.2</td>
<td>74.8</td>
<td>62.1</td>
<td>59.4</td>
<td>4%</td>
</tr>
<tr>
<td></td>
<td>ImageNet</td>
<td></td>
<td>43.9</td>
<td>31.8</td>
<td>46.9</td>
<td>60.7</td>
<td>59.3</td>
<td>67.8</td>
<td>60.7</td>
<td>49.9</td>
<td>18%</td>
</tr>
<tr>
<td></td>
<td>Caltech101</td>
<td></td>
<td>20.2</td>
<td>18.8</td>
<td>11.0</td>
<td>31.4</td>
<td>100</td>
<td>29.3</td>
<td>100</td>
<td>22.2</td>
<td>78%</td>
</tr>
<tr>
<td></td>
<td>MSRC</td>
<td></td>
<td>28.6</td>
<td>17.1</td>
<td>32.3</td>
<td>21.5</td>
<td>67.7</td>
<td>74.3</td>
<td>74.3</td>
<td>33.4</td>
<td>55%</td>
</tr>
<tr>
<td>Mean others</td>
<td></td>
<td></td>
<td>42.0</td>
<td>34.7</td>
<td>34.6</td>
<td>38.2</td>
<td>57.9</td>
<td>61.7</td>
<td>72.4</td>
<td>44.8</td>
<td>48%</td>
</tr>
<tr>
<td>“person”</td>
<td>SUN09</td>
<td></td>
<td>16.1</td>
<td>11.8</td>
<td>14.0</td>
<td>7.9</td>
<td>6.8</td>
<td>23.5</td>
<td>16.1</td>
<td>12.8</td>
<td>20%</td>
</tr>
<tr>
<td></td>
<td>LabelMe</td>
<td></td>
<td>11.0</td>
<td>26.6</td>
<td>7.5</td>
<td>6.3</td>
<td>8.4</td>
<td>24.3</td>
<td>26.6</td>
<td>11.5</td>
<td>57%</td>
</tr>
<tr>
<td></td>
<td>PASCAL</td>
<td></td>
<td>11.9</td>
<td>11.1</td>
<td>20.7</td>
<td>13.6</td>
<td>48.3</td>
<td>50.5</td>
<td>20.7</td>
<td>27.1</td>
<td>-31%</td>
</tr>
<tr>
<td></td>
<td>ImageNet</td>
<td></td>
<td>8.9</td>
<td>11.1</td>
<td>11.8</td>
<td>20.7</td>
<td>76.7</td>
<td>61.0</td>
<td>20.7</td>
<td>33.9</td>
<td>-63%</td>
</tr>
<tr>
<td></td>
<td>Caltech101</td>
<td></td>
<td>7.6</td>
<td>11.8</td>
<td>17.3</td>
<td>22.5</td>
<td>99.6</td>
<td>65.8</td>
<td>99.6</td>
<td>25.0</td>
<td>75%</td>
</tr>
<tr>
<td></td>
<td>MSRC</td>
<td></td>
<td>9.4</td>
<td>15.5</td>
<td>15.3</td>
<td>15.3</td>
<td>93.4</td>
<td>78.4</td>
<td>78.4</td>
<td>29.8</td>
<td>62%</td>
</tr>
<tr>
<td>Mean others</td>
<td></td>
<td></td>
<td>9.8</td>
<td>12.3</td>
<td>13.2</td>
<td>13.1</td>
<td>46.7</td>
<td>45.0</td>
<td>43.7</td>
<td>23.4</td>
<td>47%</td>
</tr>
</tbody>
</table>

Mean others
# Negative Set Bias

Table 2. Measuring Negative Set Bias.

<table>
<thead>
<tr>
<th>task</th>
<th>Negative Set:</th>
<th>Positive Set:</th>
<th>SUN09</th>
<th>LabelMe</th>
<th>PASCAL</th>
<th>ImageNet</th>
<th>Caltech101</th>
<th>MSRC</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>“car” detection</td>
<td>self</td>
<td></td>
<td>67.6</td>
<td>62.4</td>
<td>56.3</td>
<td>60.5</td>
<td>97.7</td>
<td>74.5</td>
<td>70.0</td>
</tr>
<tr>
<td></td>
<td>all</td>
<td></td>
<td>53.8</td>
<td>51.3</td>
<td>47.1</td>
<td>65.2</td>
<td>97.7</td>
<td>70.0</td>
<td>64.1</td>
</tr>
<tr>
<td></td>
<td>percent drop</td>
<td></td>
<td>20%</td>
<td>18%</td>
<td>16%</td>
<td>-8%</td>
<td>0%</td>
<td>6%</td>
<td>8%</td>
</tr>
<tr>
<td>“person” detection</td>
<td>self</td>
<td></td>
<td>67.4</td>
<td>68.6</td>
<td>53.8</td>
<td>60.4</td>
<td>100</td>
<td>76.7</td>
<td>71.1</td>
</tr>
<tr>
<td></td>
<td>all</td>
<td></td>
<td>52.2</td>
<td>58.0</td>
<td>42.6</td>
<td>63.4</td>
<td>100</td>
<td>71.5</td>
<td>64.6</td>
</tr>
<tr>
<td></td>
<td>percent drop</td>
<td></td>
<td>22%</td>
<td>15%</td>
<td>21%</td>
<td>-5%</td>
<td>0%</td>
<td>7%</td>
<td>9%</td>
</tr>
</tbody>
</table>
Table 3. “Market Value” for a “car” sample across datasets

<table>
<thead>
<tr>
<th></th>
<th>SUN09 market</th>
<th>LabelMe market</th>
<th>PASCAL market</th>
<th>ImageNet market</th>
<th>Caltech101 market</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 SUN09 is worth</td>
<td>1 SUN09</td>
<td>0.91 LabelMe</td>
<td>0.72 pascal</td>
<td>0.41 ImageNet</td>
<td>0 Caltech</td>
</tr>
<tr>
<td>1 LabelMe is worth</td>
<td>0.41 SUN09</td>
<td>1 LabelMe</td>
<td>0.26 pascal</td>
<td>0.31 ImageNet</td>
<td>0 Caltech</td>
</tr>
<tr>
<td>1 pascal is worth</td>
<td>0.29 SUN09</td>
<td>0.50 LabelMe</td>
<td>1 pascal</td>
<td>0.88 ImageNet</td>
<td>0 Caltech</td>
</tr>
<tr>
<td>1 ImageNet is worth</td>
<td>0.17 SUN09</td>
<td>0.24 LabelMe</td>
<td>0.40 pascal</td>
<td>1 ImageNet</td>
<td>0 Caltech</td>
</tr>
<tr>
<td>1 Caltech101 is worth</td>
<td>0.18 SUN09</td>
<td>0.23 LabelMe</td>
<td>0 pascal</td>
<td>0.28 ImageNet</td>
<td>1 Caltech</td>
</tr>
<tr>
<td>Basket of Currencies</td>
<td>0.41 SUN09</td>
<td>0.58 LabelMe</td>
<td>0.48 pascal</td>
<td>0.58 ImageNet</td>
<td>0.20 Caltech</td>
</tr>
</tbody>
</table>
Overall…

- Caltech101, MSRC – bad
- PASCAL, ImageNet -- better
2. We will never have enough data
Long Tails -- Unfamiliar is Common

10% of the objects account for 90% of the data

~Zipf's law
Dealing with sparse data (rare scenes)

Quick Fixes:
better alignment
  • e.g. reduce resolution, sifting, warping, etc.
segment into chunks
  • e.g. segmentation for recognition approaches

Understand the simple stuff first
Recognize when it’s easy!

People take on a variety of poses, aspects, scales

- self-occlusion
- rare pose
- motion blur
- non-distinctive pose
- too small
- just right detect this

Ramanan, Forsyth, Zisserman, 2004
"Poping out" foreground objects

Hoiem et al, ICCV 2007

Figure 10. Object popout. We show five out of the fifteen most "solid" regions in the Geometric Context dataset. Our algorithm often finds foreground objects, which would be helpful for unsupervised object discovery [24].
Guess structure

David C. Lee, Martial Hebert, Takeo Kanade, CVPR’09
Guess structure

David C. Lee, Martial Hebert, Takeo Kanade, CVPR’09
Subtracting away structure

Structure

Objects

Wall appearance modeling

David C. Lee, Martial Hebert, Takeo Kanade, CVPR’09
Dealing with sparse data (rare scenes)

Long-term Fixes:
Attributes – densifying the labels
From categorization to association
  • Ask not “what is this?”, ask “what is this like?”
Categorization vs. The Data
categorization is losing…

VS.

Yahoo!

vs.

Google
“…That which we call a rose
By any other name would smell as sweet.”

“chair” category (PASCAL VOC)

“train” category (PASCAL VOC)
Discriminative Exemplar-based Detector

- Train a linear SVM for each positive instance
  - with lots of mined hard-negatives
- Use leave-one-out cross-validation to calibrate detectors
- At test time, run all detectors through non-max suppression to find winner

Malisiewicz et al, ICCV’11
What’s Going On?

• Instead of one hard problem, many easy problems
• Each detector is an “associator”, an expert in it’s local neighborhood only
• More powerful than local distance learning – exemplar doesn’t have to reside at origin
• The negatives define the boundary
• Related to one-class SVMs, kernel SVMs, kernel learning, KNN-SVM… but no need for common kernel. Also get associations.
A sample instance detector
Results
Label Transfer

• Now can easily transfer labels, segmentations, layouts, even 3D models:
Take-home Message

Large-scale data is necessary, but certainly not sufficient, to solve recognition

Corollary: all the coolest stuff hasn’t been done yet!