Object Detection Wrapup
Things to remember

• Sliding window for search

• Features based on differences of intensity (gradient, wavelet, etc.)
  – Excellent results require careful feature design

• Boosting for feature selection

• Integral images, cascade for speed

• Bootstrapping to deal with many, many negative examples
Discriminative part-based models

Root filter  Part filters  Deformation weights

P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan, Object Detection with Discriminatively Trained Part Based Models, PAMI 32(9), 2010
Car model

Component 1

Component 2
Person model
Bottle model
More detections

horse

sofa

bottle
The PASCAL Visual Object Classes Challenge 2009 (VOC2009)

- Twenty object categories (aeroplane to TV/monitor)

- Three challenges:
  - Classification challenge (is there an X in this image?)
  - Detection challenge (draw a box around every X)
  - Segmentation challenge
Dataset: Collection

- Images downloaded from **flickr**
  - 500,000 images downloaded and random subset selected for annotation
Dataset: Annotation

- Complete annotation of all objects
- Annotated over web with **written guidelines**
  - High quality (?)
Examples

Aeroplane  Bicycle  Bird  Boat  Bottle

Bus  Car  Cat  Chair  Cow
<table>
<thead>
<tr>
<th>Dining Table</th>
<th>Dog</th>
<th>Horse</th>
<th>Motorbike</th>
<th>Person</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Dining Table" /></td>
<td><img src="image2" alt="Dog" /></td>
<td><img src="image3" alt="Horse" /></td>
<td><img src="image4" alt="Motorbike" /></td>
<td><img src="image5" alt="Person" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Potted Plant</th>
<th>Sheep</th>
<th>Sofa</th>
<th>Train</th>
<th>TV/Monitor</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image6" alt="Potted Plant" /></td>
<td><img src="image7" alt="Sheep" /></td>
<td><img src="image8" alt="Sofa" /></td>
<td><img src="image9" alt="Train" /></td>
<td><img src="image10" alt="TV/Monitor" /></td>
</tr>
</tbody>
</table>
Classification Challenge

- Predict whether at least one object of a given class is present in an image

is there a cat?
Evaluation

- **Average Precision [TREC]** averages precision over the entire range of recall
  - Curve interpolated to reduce influence of “outliers”

- A good score requires both high recall and high precision
- Application-independent
- Penalizes methods giving high precision but low recall
Participation

- 48 Methods, 20 Groups
### Results: AP by Method and Class

<table>
<thead>
<tr>
<th>Method</th>
<th>aero plane</th>
<th>bicycle</th>
<th>bird</th>
<th>boat</th>
<th>bottle</th>
<th>bus</th>
<th>car</th>
<th>cat</th>
<th>chair</th>
<th>cow</th>
<th>dining table</th>
<th>dog</th>
<th>horse</th>
<th>motor bike</th>
<th>person</th>
<th>potted plant</th>
<th>sheep</th>
<th>sofa</th>
<th>train</th>
<th>tv/monitor</th>
</tr>
</thead>
<tbody>
<tr>
<td>CVC_FLAT</td>
<td>85.3</td>
<td>57.8</td>
<td>66.0</td>
<td>66.1</td>
<td>36.2</td>
<td>70.6</td>
<td>60.6</td>
<td>63.5</td>
<td>55.1</td>
<td>44.6</td>
<td>53.4</td>
<td>49.1</td>
<td>64.4</td>
<td>66.8</td>
<td>84.8</td>
<td>37.4</td>
<td>44.1</td>
<td>47.9</td>
<td>81.9</td>
<td>67.5</td>
</tr>
<tr>
<td>CVC_FLAT-HOG-ESS</td>
<td>86.3</td>
<td>60.7</td>
<td>66.4</td>
<td>65.3</td>
<td>41.0</td>
<td>71.7</td>
<td>64.7</td>
<td>63.9</td>
<td>55.5</td>
<td>40.1</td>
<td>51.3</td>
<td>45.9</td>
<td>65.2</td>
<td>68.9</td>
<td>85.0</td>
<td>40.8</td>
<td>49.0</td>
<td>49.1</td>
<td>81.8</td>
<td>68.6</td>
</tr>
<tr>
<td>CVC_PLUS</td>
<td>86.6</td>
<td>58.4</td>
<td>66.7</td>
<td>67.3</td>
<td>34.8</td>
<td>70.4</td>
<td>60.0</td>
<td>64.2</td>
<td>52.5</td>
<td>43.0</td>
<td>50.8</td>
<td>46.5</td>
<td>64.1</td>
<td>66.8</td>
<td>84.4</td>
<td>37.5</td>
<td>45.1</td>
<td>45.4</td>
<td>82.1</td>
<td>67.0</td>
</tr>
<tr>
<td>FIRSTNIKON_AVGSRKDA</td>
<td>83.3</td>
<td>59.3</td>
<td>62.7</td>
<td>65.3</td>
<td>30.2</td>
<td>71.6</td>
<td>58.2</td>
<td>62.2</td>
<td>54.3</td>
<td>40.7</td>
<td>49.2</td>
<td>50.0</td>
<td>66.6</td>
<td>62.9</td>
<td>83.3</td>
<td>34.2</td>
<td>48.2</td>
<td>46.1</td>
<td>83.4</td>
<td>65.5</td>
</tr>
<tr>
<td>FIRSTNIKON_AVGSVM</td>
<td>83.8</td>
<td>58.2</td>
<td>62.6</td>
<td>65.2</td>
<td>32.0</td>
<td>69.8</td>
<td>57.7</td>
<td>61.1</td>
<td>54.5</td>
<td>44.0</td>
<td>50.3</td>
<td>49.6</td>
<td>64.6</td>
<td>61.7</td>
<td>83.2</td>
<td>33.4</td>
<td>46.5</td>
<td>48.0</td>
<td>81.6</td>
<td>65.3</td>
</tr>
<tr>
<td>FIRSTNIKON_BOOSTSRKDA</td>
<td>83.0</td>
<td>59.2</td>
<td>61.4</td>
<td>64.6</td>
<td>33.2</td>
<td>71.1</td>
<td>57.5</td>
<td>61.0</td>
<td>54.8</td>
<td>40.7</td>
<td>48.3</td>
<td>50.0</td>
<td>65.5</td>
<td>63.4</td>
<td>82.8</td>
<td>32.8</td>
<td>47.0</td>
<td>47.1</td>
<td>83.3</td>
<td>64.6</td>
</tr>
<tr>
<td>FIRSTNIKON_BOOSTSVM</td>
<td>83.5</td>
<td>56.8</td>
<td>61.8</td>
<td>65.5</td>
<td>33.2</td>
<td>69.7</td>
<td>57.3</td>
<td>60.5</td>
<td>54.6</td>
<td>43.1</td>
<td>48.3</td>
<td>50.3</td>
<td>64.3</td>
<td>62.4</td>
<td>82.3</td>
<td>32.9</td>
<td>46.9</td>
<td>48.4</td>
<td>82.0</td>
<td>64.2</td>
</tr>
<tr>
<td>LEAR_CHI-SVM-MULT-LOC</td>
<td>79.5</td>
<td>55.5</td>
<td>54.5</td>
<td>63.9</td>
<td>43.7</td>
<td>70.3</td>
<td>66.4</td>
<td>56.5</td>
<td>54.4</td>
<td>38.8</td>
<td>44.1</td>
<td>46.2</td>
<td>58.5</td>
<td>64.2</td>
<td>82.2</td>
<td>39.1</td>
<td>41.3</td>
<td>39.8</td>
<td>73.6</td>
<td>66.2</td>
</tr>
<tr>
<td>NECUIUC_CDCV</td>
<td>88.1</td>
<td>68.0</td>
<td>68.0</td>
<td>72.5</td>
<td>41.0</td>
<td>78.9</td>
<td>70.4</td>
<td>70.4</td>
<td>58.1</td>
<td>53.4</td>
<td>55.7</td>
<td>59.3</td>
<td>73.1</td>
<td>71.3</td>
<td>84.5</td>
<td>32.3</td>
<td>53.3</td>
<td>56.7</td>
<td>86.0</td>
<td>66.8</td>
</tr>
<tr>
<td>NECUIUC_CLS-DTCT</td>
<td>88.0</td>
<td>68.6</td>
<td>67.9</td>
<td>72.9</td>
<td>44.2</td>
<td>79.5</td>
<td>72.5</td>
<td>70.8</td>
<td>59.5</td>
<td>53.6</td>
<td>57.5</td>
<td>59.0</td>
<td>72.6</td>
<td>72.3</td>
<td>85.3</td>
<td>36.6</td>
<td>56.9</td>
<td>57.9</td>
<td>85.9</td>
<td>68.0</td>
</tr>
<tr>
<td>NECUIUC_LL-CDCV</td>
<td>87.1</td>
<td>67.4</td>
<td>65.8</td>
<td>72.3</td>
<td>40.9</td>
<td>78.3</td>
<td>69.7</td>
<td>69.7</td>
<td>58.5</td>
<td>50.1</td>
<td>55.1</td>
<td>56.3</td>
<td>71.8</td>
<td>70.8</td>
<td>84.1</td>
<td>31.4</td>
<td>51.5</td>
<td>55.1</td>
<td>84.7</td>
<td>65.2</td>
</tr>
<tr>
<td>NECUIUC_LN-CDCV</td>
<td>87.7</td>
<td>67.8</td>
<td>68.1</td>
<td>71.1</td>
<td>39.1</td>
<td>78.5</td>
<td>70.6</td>
<td>70.7</td>
<td>57.4</td>
<td>51.7</td>
<td>53.3</td>
<td>59.2</td>
<td>71.6</td>
<td>70.6</td>
<td>84.0</td>
<td>30.9</td>
<td>51.7</td>
<td>55.9</td>
<td>85.9</td>
<td>66.7</td>
</tr>
<tr>
<td>UVASURREY_BASELINE</td>
<td>84.1</td>
<td>59.2</td>
<td>62.7</td>
<td>65.4</td>
<td>35.7</td>
<td>70.6</td>
<td>59.8</td>
<td>61.3</td>
<td>56.7</td>
<td>45.3</td>
<td>52.4</td>
<td>50.6</td>
<td>66.1</td>
<td>66.6</td>
<td>83.7</td>
<td>34.8</td>
<td>47.2</td>
<td>47.7</td>
<td>80.8</td>
<td>65.9</td>
</tr>
<tr>
<td>UVASURREY_MKFDA+BOW</td>
<td>84.7</td>
<td>63.9</td>
<td>66.1</td>
<td>67.3</td>
<td>37.9</td>
<td>74.1</td>
<td>63.2</td>
<td>64.0</td>
<td>57.1</td>
<td>46.2</td>
<td>54.7</td>
<td>53.5</td>
<td>68.1</td>
<td>70.6</td>
<td>85.2</td>
<td>38.5</td>
<td>47.2</td>
<td>49.3</td>
<td>83.2</td>
<td>68.1</td>
</tr>
<tr>
<td>UVASURREY_TUNECOLORKERNELSEL</td>
<td>85.0</td>
<td>62.8</td>
<td>65.1</td>
<td>66.5</td>
<td>37.6</td>
<td>73.5</td>
<td>62.1</td>
<td>62.0</td>
<td>57.4</td>
<td>45.1</td>
<td>54.5</td>
<td>52.5</td>
<td>67.7</td>
<td>69.8</td>
<td>84.8</td>
<td>39.1</td>
<td>46.8</td>
<td>49.9</td>
<td>82.9</td>
<td>68.1</td>
</tr>
<tr>
<td>UVASURREY_TUNECOLORSPECCKDA</td>
<td>84.6</td>
<td>62.4</td>
<td>65.6</td>
<td>67.2</td>
<td>39.4</td>
<td>74.0</td>
<td>63.4</td>
<td>62.8</td>
<td>56.7</td>
<td>43.8</td>
<td>54.7</td>
<td>52.7</td>
<td>67.3</td>
<td>70.6</td>
<td>85.0</td>
<td>38.8</td>
<td>46.9</td>
<td>50.0</td>
<td>82.2</td>
<td>66.2</td>
</tr>
</tbody>
</table>

- Only methods in 1st, 2nd or 3rd place by group shown
- Groups: CVC, FIRST/Nikon, NEC/UIUC, UVA/Surrey
Max AP: 88.1% (aeroplane) ... 40.8% (potted plant)
Precision/Recall: Aeroplane (All)
Precision/Recall: Potted plant (Top 10 by AP)
Ranked Images: Aeroplane

- Class images:
  Highest ranked
 Ranked Images: Chair

- Class images: Highest ranked
Detection Challenge

- Predict the bounding boxes of all objects of a given class in an image (if any)
Evaluating Bounding Boxes

- **Area of Overlap (AO) Measure**

\[ AO(B_{gt}, B_{p}) = \frac{|B_{gt} \cap B_{p}|}{|B_{gt} \cup B_{p}|} \]

- Need to define a threshold \( t \) such that \( AO(B_{gt}, B_{p}) \) implies a correct detection: 50\%
AP by Class

Chance essentially 0
Precision/Recall - Aeroplane
Precision/Recall – Potted plant
True Positives - Person

UoCTTI_LSVM-MDPM

MIZZOU_DEF-HOG-LBP

NECUIUC_CLS-DTCT
False Positives - Person

UoCTTI_LSVM-MDPM

MIZZOU_DEF-HOG-LBP

NECUIUC_CLS-DTCT
“Near Misses” - Person

UoCTTI_LSVM-MDPM

MIZZOU_DEF-HOG-LBP

NECUIUC_CLS-DTCT
True Positives - Bicycle

UoCTTI_LSV-MDPM

OXFORD_MKL

NECUIUC_CLS-DTCT
False Positives - Bicycle

UoCTTI_LSVM-MDPM

OXFORD_MKL

NECUUIUC_CLS-DTCT
Opportunities of Scale

Computer Vision

James Hays

Many slides from James Hays, Alyosha Efros, and Derek Hoiem

Graphic from Antonio Torralba
Computer Vision so far

• The geometry of image formation
  – Ancient / Renaissance
• Signal processing / Convolution
  – 1800, but really the 50’s and 60’s
• Hand-designed Features for recognition, either instance-level or categorical
  – 1999 (SIFT), 2003 (Video Google), 2005 (Dalal-Triggs), 2006 (spatial pyramid)
• Learning from Data
  – 1991 (EigenFaces) but late 90’s to now especially
What has changed in the last decade?

- The Internet
- Crowdsourcing
- Learning representations from the data these sources provide (deep learning)
Opportunities of Scale: Data-driven methods

- Scene completion (Today’s class)
- Im2gps
Google and massive data-driven algorithms

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
The Unreasonable Effectiveness of Data

Peter Norvig
Google

If a machine can convincingly simulate an intelligent conversation, does it necessarily understand? In the experiment, Searle imagines himself in a room, acting as a computer by manually executing a program that convincingly simulates the behavior of a native Chinese speaker.

Most of the discussion consists of attempts to refute it. "The overwhelming majority," notes BBS editor Stevan Harnad, "still think that the Chinese Room Argument is dead wrong." The sheer volume of the literature that has grown up around it inspired Pat Hayes to quip that the field of cognitive science ought to be redefined as "the ongoing research program of showing Searle's Chinese Room Argument to be false."
Questions from the piece:

Q1. Does the Chinese Room argument prove the impossibility of machine consciousness?
A1: Hell no. ... See More

Can Machines Become Moral?
The question is heard more and more often, both from those who think that machines cannot become moral, and who think that to believe otherwise is a dangerous illusion, and from those who think that machines must become moral....

BIGQUESTIONSONLINE.COM | BY DON HOWARD
Big Idea

• Do we need computer vision systems to have strong AI-like reasoning about our world?
• What if invariance / generalization isn’t actually the core difficulty of computer vision?
• What if we can perform high level reasoning with brute-force, data-driven algorithms?
Image Completion Example

[Hays and Efros. Scene Completion Using Millions of Photographs. SIGGRAPH 2007 and CACM October 2008.]

http://graphics.cs.cmu.edu/projects/scene-completion/
What should the missing region contain?
Which is the original?

(a)

(b)

(c)
How it works

• Find a similar image from a large dataset
• Blend a region from that image into the hole
Hopefully, If you have enough images, the dataset will contain very similar images that you can find with simple matching methods.
How many images is enough?
Nearest neighbors from a collection of 20 thousand images
Nearest neighbors from a collection of 2 million images
Image Data on the Internet

• Flickr (as of Sept. 19th, 2010)
  – 5 billion photographs
  – 100+ million geotagged images

• Facebook (as of 2009)
  – 15 billion

Image Data on the Internet

• Flickr (as of Nov 2013)
  – 10 billion photographs
  – 100+ million geotagged images
  – 3.5 million a day

• Facebook (as of Sept 2013)
  – 250 billion+
  – 300 million a day

• Instagram
  – 55 million a day
Image completion: how it works

[Hays and Efros. Scene Completion Using Millions of Photographs. SIGGRAPH 2007 and CACM October 2008.]
The Algorithm
Scene Matching
Scene Descriptor
Scene Descriptor

Scene Gist Descriptor
(Oliva and Torralba 2001)
Scene Descriptor

Scene Gist Descriptor

(Oliva and Torralba 2001)
2 Million Flickr Images
Context Matching
Graph cut + Poisson blending
Result Ranking

We assign each of the 200 results a score which is the sum of:

The scene matching distance

The context matching distance (color + texture)

The graph cut cost
... 200 scene matches
Which is the original?