CSE 590V: Computer vision seminar

Fall 2011

Course webpage: http://www.cs.washington.edu/education/courses/cse590v/11au/
Logistics

• Time: Tuesdays from 1:30pm-2:30pm
• Location: CSE 403
• Organizers: Neeraj Kumar and Bryan Russell
• Class mailing list: cse590v @ cs washington edu (subscribe at course webpage)
CSE 590V: Computer vision seminar
Fall 2011

Course description
CSE 590V is a seminar/reading group focused on recent work in computer vision. We will cover papers from recent and upcoming conferences related to computer vision (CVPR, ICCV, ECCV, NIPS, SIGGRAPH). The seminar is open to everyone. We especially encourage first year graduate students who may be considering research in computer vision or related areas to participate.

Logistics
Time: Tuesdays from 1:30pm-2:30pm
Location: CSE 403
Course description

• This is a seminar on recent work in computer vision
• We will cover papers from recent conferences related to computer vision: CVPR, ICCV, ECCV, NIPS, SIGGRAPH
• We have organized the papers into topics
• Each week, we will discuss the papers for a topic
Potential list of topics covered in class

- Datasets and active learning *(covered today)*
- Attributes *(covered next time)*
- Poselets
- Person detection
- Scene understanding
- Large scale recognition
- Learning
- Events and actions
- Language

- Cross-domain/multi-modal learning & matching
- Crowds & videos/social networks
- Shading and lighting
- Multi-view geometry
- RGB-D perception
- Cognitive science & saliency
- Misc/cool papers
Course expectations (everybody)

• Read the assigned paper(s) beforehand
• Come ready to discuss the papers
• Make a list of 3 items to discuss, for example:
  – Question
  – Extension
  – Critique
Course expectations (students)

• Give a presentation on one of the topics
• E-mail Neeraj or Bryan top 3 preferred topic choices by this Friday
• We will assign topics by next week
• Friday before you give your presentation, meet with Neeraj and Bryan to discuss upcoming presentation
• We will award a prize to the best presentation
Volunteer(s) for next week?

• Topic: attributes
  – Attribute Learning in Large-scale Datasets. O. Russakovsky and L. Fei-Fei. Workshop on Parts and Attributes, assoc. with ECCV 2010.
Summary to do list

• Today: sign up for course mailing list
• Friday: e-mail Neeraj and Bryan top 3 preferred topics
• Next Tuesday: read assigned attribute papers (we will let you know which ones we will focus on)
Datasets for object recognition and scene understanding

Bryan Russell
In the beginning, things weren’t always so easy...
Figure 4. This figure provides a high quality reproduction of the six images discussed in the text. (a) and (b) were taken with a considerably modified Information International Incorporated Vidissector, and the rest were taken with a Telecama TMC-2100 vidicon camera attached to a Spatial Data Systems digitizer (Camera Eye 108). The full dynamic range from black to white is represented by 256 grey-levels. The images reproduced here were created by an Optronics P1500 Photowriter from intensity arrays that measured 128 elements square. This size of intensity array corresponds to viewing a 1 in square at 5 ft with the human retina. The image of the period at the end of this sentence probably covers more than 40 retinal receptors. The reader should view the images from a distance of about 5 ft when assessing the performance of the programs.
FIGURE 4. This figure provides a high quality reproduction of the six images discussed in the text. (a) and (b) were taken with a considerably modified Information International Incorporated Vidissector, and the rest were taken with a Telemation TMC-2100 vidicon camera attached to a Spatial Data Systems digitizer (Camera Eye 108). The full dynamic range from black to white is represented by 256 grey-levels. The images reproduced here were created by an Optronics P1500hPhotowriter from intensity arrays that measured 128 elements square. This size of intensity array corresponds to viewing a 1 in square at 5 ft with the human retina. The image of the period at the end of this sentence probably covers more than 40 retinal receptors. The reader should view the images from a distance of about 5 ft when assessing the performance of the programs.
The rise of the modern dataset...
Caltech 101 and 256

101 object classes

101 object classes

256 object classes

Fei-Fei, Fergus, Perona, 2004

9,146 images

256 object classes

Griffin, Holub, Perona, 2007

30,607 images

Slide credit: A. Torralba
Table of results for Caltech 101

This is a table documenting some of the best results some paper obtained in Caltech-101 dataset.

Results shown here are all trained using 30 samples from each category.

1. **Group-Sensitive Multiple Kernel Learning for Object Categorization** (ICCV 2009)
   - Cited 17 times. 84.3%
   - Additional Info: GS-MKL

2. **LP-Beta + Geometric blur + PHOW gray/color + Self-Similarity**
   - 82.1% ± 0.3%

3. **Learning Subcategory Relevances for Category Recognition** (CVPR 2008)
   - Cited 19 times. 81.9%

4. **Object Recognition as Ranking Holistic Figure-Ground Hypotheses** (CVPR 2010)
   - Cited 8 times. 81.9%
   - Additional Info: Regression with Post-Processing.

5. **Image Classification using Random Forests and Ferns** (2007)
   - Cited 130 times. 81.3%
   - Additional Info: Bosch Multi-way SVM

6. **In Defense of Nearest-Neighbor Based Image Classification** (CVPR 2008)
   - Cited 139 times. 79.23%
   - Additional Info: NBNN (5 descriptors)

7. **Visual Geometric Group (VGG)'s Implementation of Multiple Kernel Image Classifier trained on dense SIFT, self-similarity, and geometric blur features**
   - 78.20% ± 0.4%
   - Additional Info: Result of 77.8% is obtained by combining dense SIFT, self-similarity, and geometric blur features with the multiple kernel learning
591 images, 23 object classes
Pixel-wise segmentation

J. Winn, A. Criminisi, and T. Minka, 2005
LabelMe

Tool went online July 1st, 2005
825,597 object annotations collected
199,250 images available for labeling

labelme.csail.mit.edu
Quality of the labeling

- Car: 8, 15, 22, 25%
- Person: 7, 12, 21, 25%
- Dog: 16, 28, 52, 50%
- Bird: 13, 37, 168, 75%
- Chair: 7, 10, 15, 25%
- Street lamp: 5, 9, 15, 50%
- Motorbike: 12, 22, 36, 75%
- Boat: 6, 9, 15, 25%
- Tree: 11, 20, 36, 50%
- Mug: 6, 8, 11, 75%
- Bottle: 7, 8, 11, 25%
- House: 5, 7, 12, 50%

Average labeling quality:

25%  50%  75%
Extreme labeling
The other extreme of extreme labeling

... things do not always look good...
Testing

Most common labels:
test
adksdsda
woiieiiie
...
Sophisticated testing

Most common labels:
Star
Square
Nothing
...
Creative testing
Do not try this at home

Most common labels:
Stupid birdie
Tourist checking hottie
Man’s derriere
…
2011 version - 20 object classes:

*Person*: person
*Animal*: bird, cat, cow, dog, horse, sheep
*Vehicle*: aeroplane, bicycle, boat, bus, car, motorbike, train
*Indoor*: bottle, chair, dining table, potted plant, sofa, tv/monitor

The train/val data has 11,530 images containing 27,450 ROI annotated objects and 5,034 segmentations

- Three main competitions: classification, detection, and segmentation
- Three "taster" competitions: person layout, action classification, and ImageNet large scale recognition

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, A. Zisserman
### Classification Results: VOC2010 data

Competition "comp1" (train on VOC2010 data)

#### Average Precision (AP %)

<table>
<thead>
<tr>
<th>Method</th>
<th>AP%</th>
</tr>
</thead>
<tbody>
<tr>
<td>BONN FGT SEG</td>
<td>88.0</td>
</tr>
<tr>
<td>BUPT LPBETA_MULTIFEAT</td>
<td>82.1</td>
</tr>
<tr>
<td>BUPT SVM SC HOG</td>
<td>79.6</td>
</tr>
<tr>
<td>BUPT SVM_MULTIFEAT</td>
<td>81.1</td>
</tr>
<tr>
<td>BUT FU SVM SIFT</td>
<td>89.7</td>
</tr>
<tr>
<td>CVC FLAT</td>
<td>89.4</td>
</tr>
<tr>
<td>CVC_PLUS</td>
<td>91.0</td>
</tr>
<tr>
<td>CVC_PLUSDET</td>
<td>91.7</td>
</tr>
<tr>
<td>HIT PROTOLEARN 2</td>
<td>60.7</td>
</tr>
<tr>
<td>LIG MSVM_FUSE CONCEPT</td>
<td>74.4</td>
</tr>
<tr>
<td>LIPSUPMC_KSVMBASELINE</td>
<td>78.4</td>
</tr>
<tr>
<td>LIPSUPMC_MKL_L1</td>
<td>78.5</td>
</tr>
<tr>
<td>LIPSUPMC_RANKING</td>
<td>78.6</td>
</tr>
<tr>
<td>LHIS_MKL_TRAINVAL</td>
<td>87.5</td>
</tr>
<tr>
<td>NEC_V1_HOGLBP_NONLIN_SVM</td>
<td>93.3</td>
</tr>
<tr>
<td>NEC_V1_HOGLBP_NONLIN_SVMDET</td>
<td>93.3</td>
</tr>
<tr>
<td>NII SVMSIFT</td>
<td>69.3</td>
</tr>
<tr>
<td>NLPR VSTAR_CLS_DICTLEARN</td>
<td>90.3</td>
</tr>
<tr>
<td>NTHU LINSPARSE</td>
<td>77.9</td>
</tr>
<tr>
<td>NUDT SVM_LDP_SIFT_PKM_SPKM</td>
<td>86.1</td>
</tr>
<tr>
<td>NUDT_SVM_WHGO_SIFT_CENTRIST_LLM</td>
<td>83.5</td>
</tr>
<tr>
<td>NUSPSSL_EXCLASSIFIER</td>
<td>91.3</td>
</tr>
<tr>
<td>NUSPSSL_KERNELREFUSING</td>
<td>93.0</td>
</tr>
<tr>
<td>NUSPSSL_MEFDTSVM</td>
<td>91.9</td>
</tr>
<tr>
<td>RITSU CBVR_WKF</td>
<td>85.6</td>
</tr>
<tr>
<td>SURREY MK_KDA</td>
<td>90.6</td>
</tr>
<tr>
<td>TIT_SIFT_GMM_MKL</td>
<td>87.2</td>
</tr>
<tr>
<td>UC3M GENDISC</td>
<td>85.5</td>
</tr>
<tr>
<td>UVA BW NEWCOLOURSIFT</td>
<td>91.5</td>
</tr>
<tr>
<td>UVA BW NEWCOLOURSIFT_SRKDA</td>
<td>90.6</td>
</tr>
<tr>
<td>WLU SPIN_EMDIST</td>
<td>75.8</td>
</tr>
<tr>
<td>XRCE IFV</td>
<td>87.1</td>
</tr>
</tbody>
</table>
80,000,000 tiny images

75,000 non-abstract nouns from WordNet

7 Online image search engines

And after 1 year downloading images

Google: 80 million images

A. Torralba, R. Fergus, W.T. Freeman. PAMI 2008
• An ontology of images based on WordNet
• ImageNet currently has
  – 13,000+ categories of visual concepts
  – 10 million human-cleaned images (~700im/categ)
  – 1/3+ is released online @ www.image-net.org

Deng, Dong, Socher, Li & Fei-Fei, CVPR 2009
• Collected all the terms from WordNet that described scenes, places, and environments
  • Any concrete noun which could reasonably complete the phrase “I am in a place”, or “let’s go to the place”

• 899 scene categories
• 130,519 images
• 397 scene categories with at least 100 images
• 63,726 labeled objects
Collecting datasets
towards $10^6$-$7$ examples

- ESP game (CMU)
  Luis Von Ahn and Laura Dabbish 2004

- LabelMe (MIT)
  Russell, Torralba, Freeman, 2005

- StreetScenes (CBCL-MIT)
  Bileschi, Poggio, 2006

- WhatWhere (Caltech)
  Perona et al, 2007

- PASCAL challenges (2006-2011)
  M. Everingham et al.

- Lotus Hill Institute
  Song-Chun Zhu et al 2007

- ImageNet (Stanford)
  J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei

- Tiny images
  A. Torralba, R. Fergus and W.T. Freeman
Video: event and action recognition

Sangmin Oh, et al. 2011

## Video: event and action recognition

<table>
<thead>
<tr>
<th>Feature</th>
<th>KTH</th>
<th>Weizmann</th>
<th>HOHA 1</th>
<th>TRECVID</th>
<th>This Work</th>
</tr>
</thead>
<tbody>
<tr>
<td># of Event Types</td>
<td>6</td>
<td>10</td>
<td>8</td>
<td>10</td>
<td>23</td>
</tr>
<tr>
<td>Avg. # of samples per class</td>
<td>100</td>
<td>9</td>
<td>~85</td>
<td>3~1670</td>
<td>10~1500</td>
</tr>
<tr>
<td>Max. Resolution (w x h)</td>
<td>160 x 120</td>
<td>180 x 144</td>
<td>~540 x 240</td>
<td>720 x 576</td>
<td>1920 x 1080</td>
</tr>
<tr>
<td>Human Height in Pixels</td>
<td>80~100</td>
<td>60~70</td>
<td>100~1200</td>
<td>20~200</td>
<td>20~180</td>
</tr>
<tr>
<td>Human to video height ratio</td>
<td>65~85%</td>
<td>42~50%</td>
<td>50~500%</td>
<td>4~36%</td>
<td>2~20%</td>
</tr>
<tr>
<td># Scenes</td>
<td>N/A</td>
<td>N/A</td>
<td>Many</td>
<td>5</td>
<td>17</td>
</tr>
<tr>
<td>Viewpoint Type</td>
<td>Side</td>
<td>Side</td>
<td>Varying</td>
<td>5 / Varying</td>
<td>Varying</td>
</tr>
<tr>
<td>Natural Background Clutter</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Incidental Objects/Activities</td>
<td>No</td>
<td>No</td>
<td>Yes, Varying</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>End-to-end Activities</td>
<td>No</td>
<td>No</td>
<td>Yes, Varying</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Tight Bounding boxes</td>
<td>Cropped</td>
<td>Cropped</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Multiple annotations on movers</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Camera Motion</td>
<td>No</td>
<td>No</td>
<td>Varying</td>
<td>No</td>
<td>Varying</td>
</tr>
</tbody>
</table>

100 hours  29 hours

Sangmin Oh, et al. 2011
Unbiased Look at Dataset Bias

Alyosha Efros (CMU)
Antonio Torralba (MIT)

Disclaimer: no graduate students have been harmed in the production of this paper
Excesses of the “Data Revolution”

• Are we getting too obsessed with evaluation?
  – The dictatorship of the PR curve over the pixels…
  – Hard to jump out of algorithmic local minima
  – Too much value for “winning” a challenge
  – Easy to overfit over time

• There are all behavioral problems
  – Can be fixed with proper “Best Practices”
Are datasets measuring the right thing?

• In Machine Learning:
  Dataset is The World

• In Recognition
  Dataset is a representation of The World

• ML solution: domain transfer

• Vision question: Do datasets provide a good representation?
Visual Data is Inherently Biased

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

• But it’s not random samples of visual world
Google StreetView
Paris

Knopp, Sivic, Pajdla, ECCV 2010
Sampling Bias

- People like to take pictures on vacation
Photographer Bias

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

Little Leaguer

Kids with Santa

The Graduate

Newlyweds

“100 Special Moments” by Jason Salavon
Our Question

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

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>12</td>
<td></td>
</tr>
</tbody>
</table>

- Caltech 101
- Caltech 256
- MSRC
- UIUC cars
- Tiny Images
- Corel
- PASCAL 2007
- LabelMe
- COIL-100
- ImageNet
- 15 Scenes
- SUN’09
SVM plays “Name that dataset!”
SVM plays “Name that dataset!”

- 12 1-vs-all classifiers
- Standard full-image features
- 39% performance (chance is 8%)
SVM plays “Name that dataset!”
Dataset look-alikes

**ImageNet pretending to be:**

- Caltech 255 look-alikes from ImageNet
- COREL look-alikes from ImageNet
- MSRC look-alikes from ImageNet

**PASCAL VOC pretending to be:**

- 15 scenes look-a-likes 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

PASCAL cars

SUN cars

Caltech101 cars

ImageNet cars

LabelMe cars

Performance: 61%
(chance: 20%)
Where do this bias comes from?
Some bias is in the world
Some bias is in the world
Some bias comes from the way the data is collected
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” classification</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></td>
<td>Mean others</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” detection</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></td>
<td>Mean others</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” classification</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></td>
<td>Mean others</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>
<tr>
<td>“person” detection</td>
<td>SUN09</td>
<td></td>
<td>69.6</td>
<td>56.8</td>
<td>37.9</td>
<td>45.7</td>
<td>52.1</td>
<td>72.7</td>
<td>69.6</td>
<td>53.0</td>
<td>24%</td>
</tr>
<tr>
<td></td>
<td>LabelMe</td>
<td></td>
<td>58.9</td>
<td>66.6</td>
<td>38.4</td>
<td>43.1</td>
<td>57.9</td>
<td>68.9</td>
<td>66.6</td>
<td>53.4</td>
<td>20%</td>
</tr>
<tr>
<td></td>
<td>PASCAL</td>
<td></td>
<td>56.0</td>
<td>55.6</td>
<td>56.3</td>
<td>55.6</td>
<td>56.8</td>
<td>74.8</td>
<td>56.3</td>
<td>59.8</td>
<td>-6%</td>
</tr>
<tr>
<td></td>
<td>ImageNet</td>
<td></td>
<td>48.8</td>
<td>39.0</td>
<td>40.1</td>
<td>59.6</td>
<td>53.2</td>
<td>70.7</td>
<td>59.6</td>
<td>50.4</td>
<td>15%</td>
</tr>
<tr>
<td></td>
<td>Caltech101</td>
<td></td>
<td>24.6</td>
<td>18.1</td>
<td>12.4</td>
<td>26.6</td>
<td>100</td>
<td>31.6</td>
<td>100</td>
<td>22.7</td>
<td>77%</td>
</tr>
<tr>
<td></td>
<td>MSRC</td>
<td></td>
<td>33.8</td>
<td>18.2</td>
<td>30.9</td>
<td>20.8</td>
<td>69.5</td>
<td>74.7</td>
<td>74.7</td>
<td>34.6</td>
<td>54%</td>
</tr>
<tr>
<td></td>
<td>Mean others</td>
<td></td>
<td>44.4</td>
<td>37.5</td>
<td>31.9</td>
<td>38.4</td>
<td>57.9</td>
<td>63.7</td>
<td>71.1</td>
<td>45.6</td>
<td>36%</td>
</tr>
</tbody>
</table>
Figure 6. Cross-dataset generalization for “car” detection as function of training data
### Dataset Value

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>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>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>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>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>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>
Mixing datasets

Test on Caltech 101

Task: car detection
Features: HOG

Number training examples

AP

Training on Caltech 101
Adding additional data from PASCAL

Number training examples
Mixing datasets

Test on PASCAL

Number training examples

AP

Adding more PASCAL

Training on PASCAL

Adding more from LabelMe

Adding more from Caltech 101

0.1

0.2

0.3

0.4

0.5

0.6
Negative Set Bias

Not all the bias comes from the appearance of the objects we care about

<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>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>
<td>64.1</td>
</tr>
<tr>
<td></td>
<td>all</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></td>
<td></td>
</tr>
<tr>
<td></td>
<td>percent drop</td>
<td>20%</td>
<td>18%</td>
<td>16%</td>
<td>-8%</td>
<td>0%</td>
<td>6%</td>
<td></td>
<td>8%</td>
</tr>
<tr>
<td>“person” detection</td>
<td>self</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>
<td>64.6</td>
</tr>
<tr>
<td></td>
<td>all</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></td>
<td></td>
</tr>
<tr>
<td></td>
<td>percent drop</td>
<td>22%</td>
<td>15%</td>
<td>21%</td>
<td>-5%</td>
<td>0%</td>
<td>7%</td>
<td></td>
<td>9%</td>
</tr>
</tbody>
</table>
Overall…

- Caltech, MSRC – bad
- PASCAL, ImageNet -- better
Causes for Pessimism

• Our best-performing techniques just don’t work in the real world
  – E.g. try a person detector on Hollywood film

• The classifiers are inherently designed to overfit to type of data it’s trained on.
  – we just don’t have enough negative data to present this…
Causes for Optimism

• We are getting better. The new datasets are better than the old ones.

• Large dataset trend will alleviate this trend.
Summary

- Until now datasets are used to evaluate algorithms, but nobody has dared to evaluate them. Let's evaluate datasets.
Four Stages of Dataset Grief

1. Denial
   - WHAT BIAS? I AM SURE THAT MY MSRC CLASSIFIER WILL WORK ON ANY DATA!

2. Machine Learning
   - OF COURSE THERE IS BIAS! THAT’S WHY YOU MUST ALWAYS TRAIN AND TEST ON THE SAME DATASET.

3. Despair
   - RECOGNITION IS HOPELESS, IT WILL NEVER WORK. WE WILL JUST KEEP OVERFITTING TO THE NEXT DATASET…

4. Acceptance
   - BIAS IS HERE TO STAY, SO WE MUST BE VIGILANT THAT OUR ALGORITHMS DON’T GET DISTRACTED BY IT.