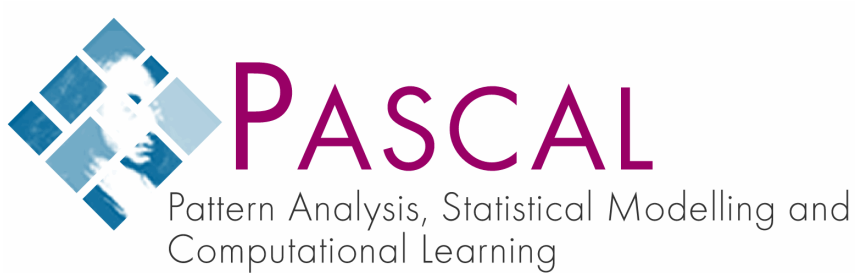
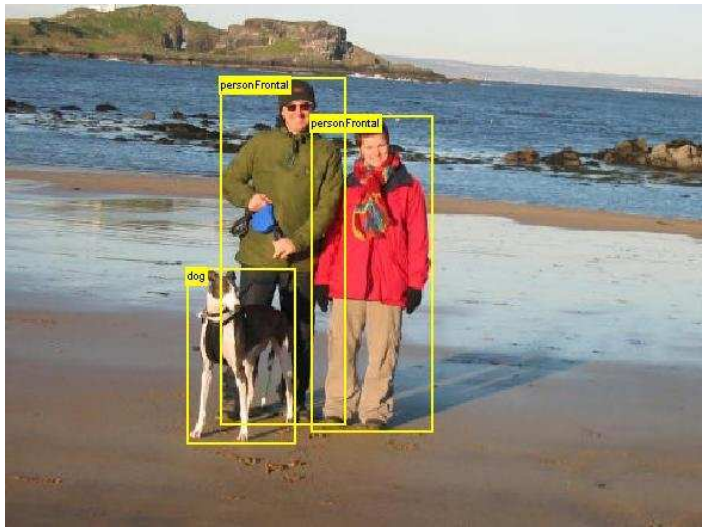
# The PASCAL Visual Object Classes (VOC) Dataset and Challenge

Mark Everingham
Luc Van Gool
Chris Williams
John Winn
Andrew Zisserman



# The PASCAL VOC Challenge

- Challenge in visual object recognition funded by PASCAL network of excellence
- Publicly available dataset of annotated images




- Main competitions in classification (is there an X in this image) and detection (where are the X's)
- "Taster competitions" in segmentation and 2-D human "pose estimation" (2007-present)

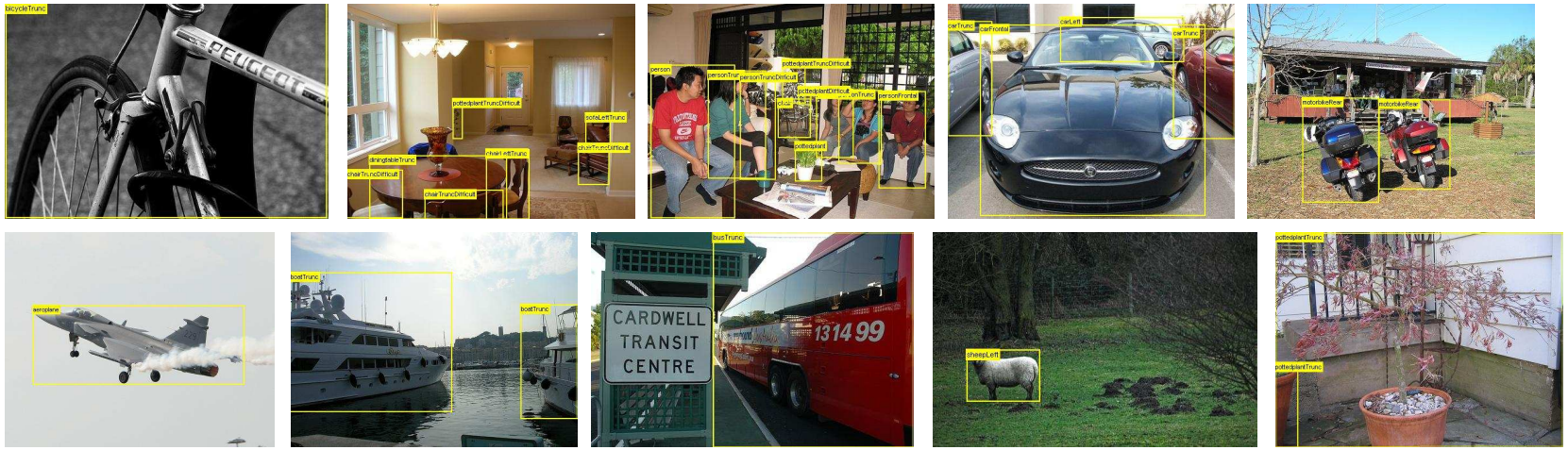
# History

| | Images | Objects | Classes | Entries | |
|---|---|---|---|---|---|
| **2005** | 2,232 | 2,871 | 4 | 12 | *Collection of existing and some new data.* |
| **2006** | 5,304 | 9,507 | 10 | 25 | *Completely new dataset from flickr (+MSRC)* |
| **2007** | 9,963 | 24,640 | 20 | 28 | *Increased classes to 20. Introduced tasters.* |
| **2008** | **8,776** | **20,739** | 20 | | *Added "occlusion" flag. Reuse of taster data. Release detailed results to support "meta-analysis"* |

- New dataset annotated annually
  - Annotation of test set is withheld until after challenge

# Dataset Content

- 20 classes: aeroplane, bicycle, boat, bottle, bus, car, cat, chair, cow, dining table, dog, horse, motorbike, person, potted plant, sheep, train, TV
- Real images not filtered for “quality” (no CC tag)



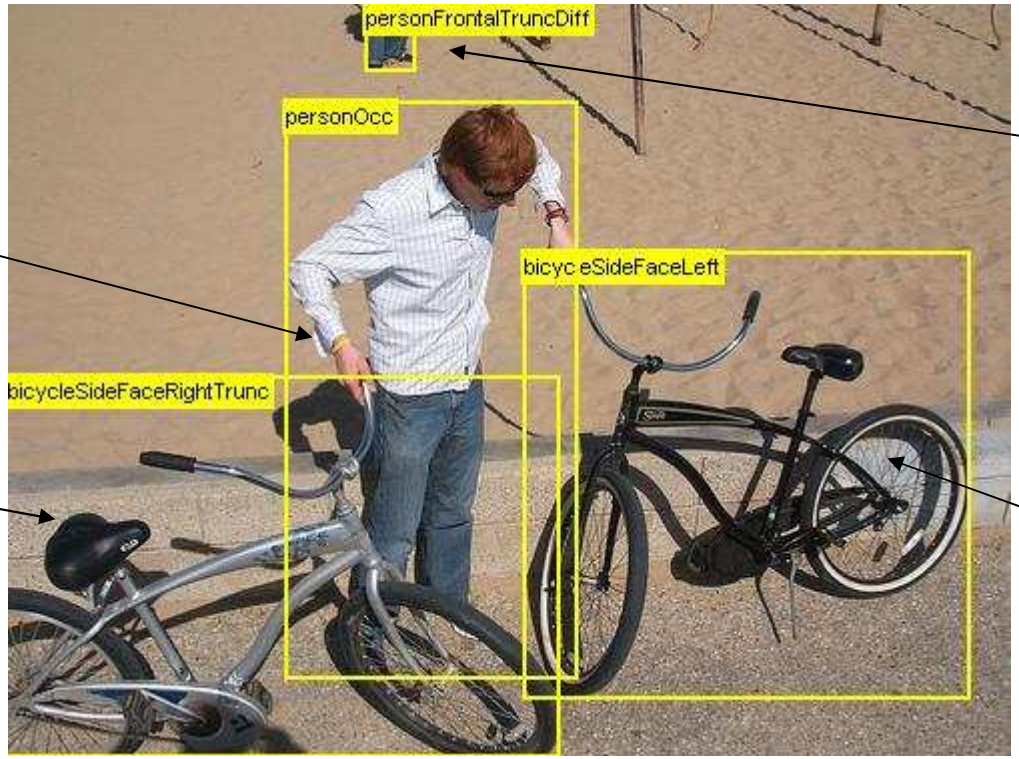

- Complex scenes, scale, pose, lighting, occlusion, ...

# Annotation

- Complete annotation of all objects
- Annotated in one session with written guidelines
  - High quality (?)

**Occluded**
Object is significantly occluded within BB

**Truncated**
Object extends beyond BB




**Difficult**
Not scored in evaluation

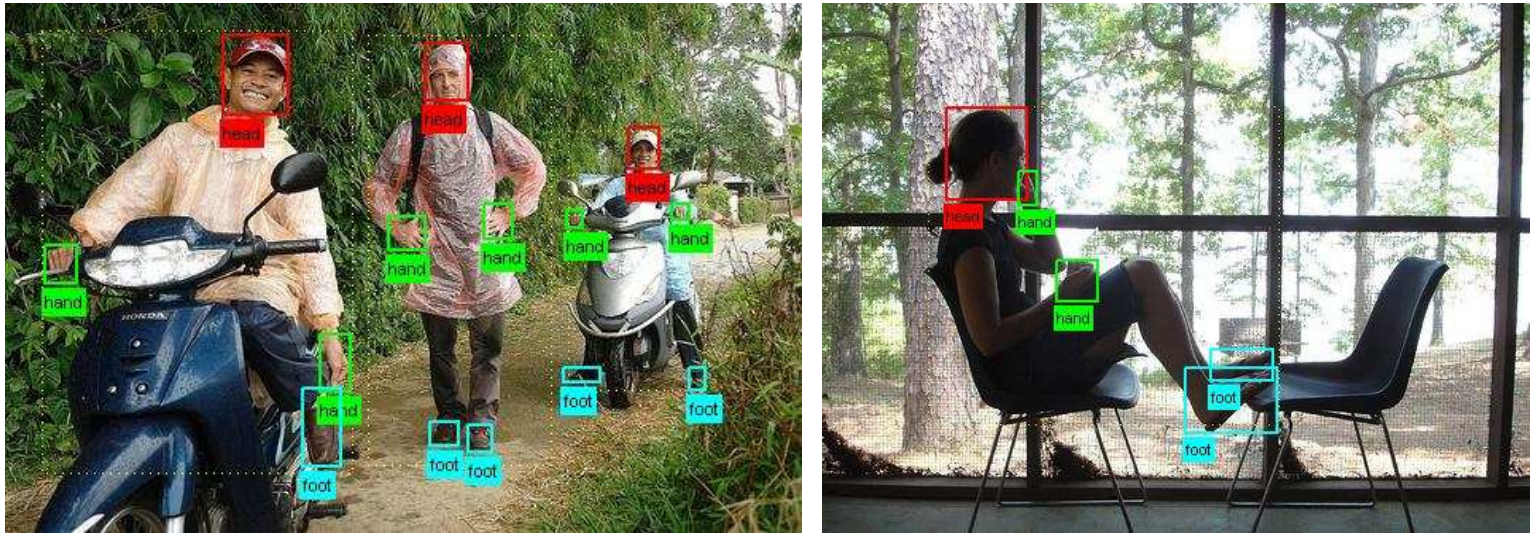
**Pose**
Facing left

# Segmentation

- Subset of images manually segmented w.r.t. 20 classes (tri-map)
  - **422** images - **1,215** objects (2007)

# 2-D “Pose” Annotation

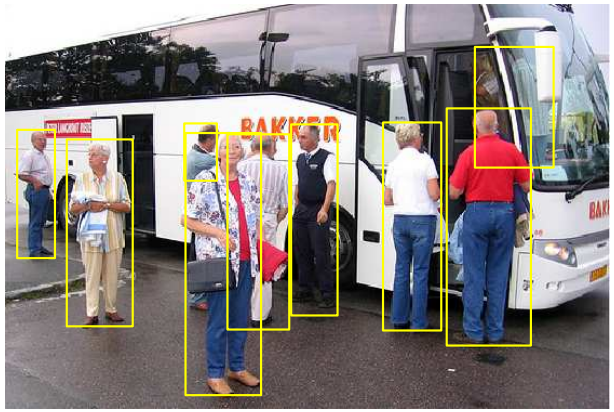
- Subset of images annotated with location of body parts – head, hands, feet
  - **322** images, **439** objects (2007)



# Main Challenge Tasks

- **Classification**
  - Is there a dog in this image?
  - Evaluation by precision/recall

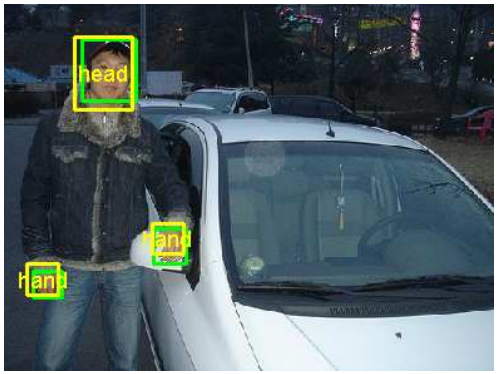
- **Detection**
  - Localize all the people (if any) in this image
  - Evaluation by precision/recall based on bounding box overlap



# “Taster” Challenges

- **“Segmentation”**
  - Label each pixel as class x or background
  - Evaluation by pixel-wise accuracy (balanced for class priors)

- **“Pose”**
  - Predict bounding boxes of body parts (2008 given bounding box of person)
  - Evaluation by precision/recall



# Attempts at Analysis

- Statistical Significance
  - Does the output of methods differ significantly?
  - Does the performance of methods differ significantly?
- What is being learnt?
  - Are confusions between classes “intuitive”?
  - Classification: learning Object or Scene?
  - Detection: is there a bias towards large objects?
- Longitudinal Results
  - Are methods getting better?

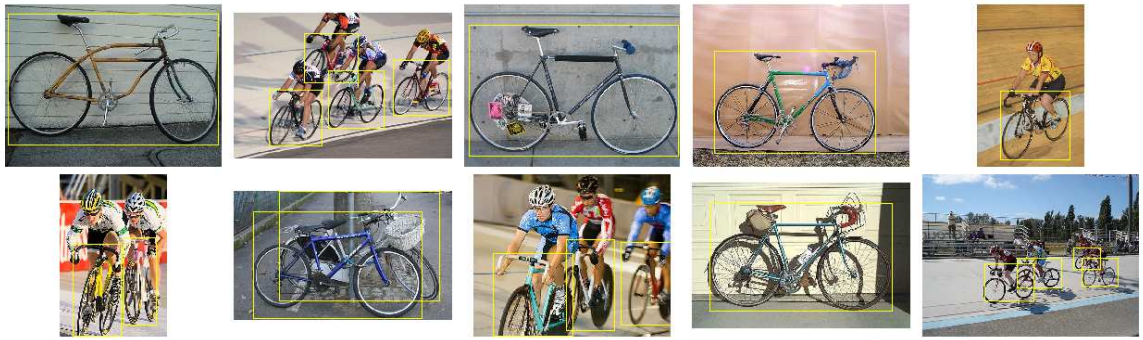
# Classification: Does output differ significantly?

- 2006: McNemar's test: Measure statistical significance of different error patterns between methods

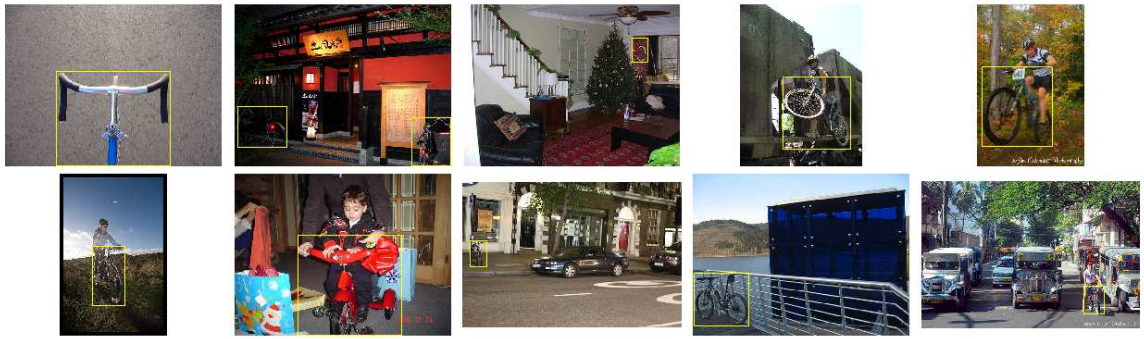
| | INRIA Nowak | QMUL HSLS | QMUL LSPCH | INRIA Marszalek | ROUND2 INRIA_Moosmann | XRCE | INRIA Moosmann | ROUND2 TKK | INRIA Larlus | UVA big5 | RWTH GMM | TKK | RWTH DiscHist | MUL 1v1 | RWTH SparseHists | MUL 1vALL | AP06 Lee | INSARouen | UVA weibull | Cambridge | Siena | AP06 Batra |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| INRIA_Nowak | - | 0.002 | 0.004 | 0.006 | 0.011 | 0.017 | 0.026 | 0.038 | 0.046 | 0.050 | 0.053 | 0.055 | 0.057 | 0.061 | 0.062 | 0.075 | 0.099 | 0.103 | 0.105 | 0.125 | 0.151 | 0.167 |
| QMUL_HSLS | -0.002 | - | 0.001 | 0.003 | 0.009 | 0.014 | 0.023 | 0.036 | 0.044 | 0.047 | 0.051 | 0.053 | 0.055 | 0.059 | 0.060 | 0.073 | 0.097 | 0.101 | 0.102 | 0.122 | 0.149 | 0.165 |
| QMUL_LSPCH | -0.004 | -0.001 | - | 0.002 | 0.007 | 0.013 | 0.022 | 0.035 | 0.042 | 0.046 | 0.050 | 0.052 | 0.054 | 0.057 | 0.059 | 0.071 | 0.096 | 0.099 | 0.101 | 0.121 | 0.147 | 0.164 |
| INRIA_Marszalek | -0.006 | -0.003 | -0.002 | - | 0.005 | 0.011 | 0.020 | 0.033 | 0.040 | 0.044 | 0.048 | 0.049 | 0.052 | 0.055 | 0.056 | 0.069 | 0.094 | 0.097 | 0.099 | 0.119 | 0.145 | 0.161 |
| ROUND2_INRIA_Moosmann | -0.011 | -0.009 | -0.007 | -0.005 | - | 0.006 | 0.015 | 0.027 | 0.035 | 0.039 | 0.042 | 0.044 | 0.046 | 0.050 | 0.051 | 0.064 | 0.088 | 0.092 | 0.094 | 0.114 | 0.140 | 0.156 |
| XRCE | -0.017 | -0.014 | -0.013 | -0.011 | -0.006 | - | 0.009 | 0.022 | 0.029 | 0.033 | 0.037 | 0.039 | 0.041 | 0.044 | 0.046 | 0.058 | 0.083 | 0.086 | 0.088 | 0.108 | 0.134 | 0.151 |
| INRIA_Moosmann | -0.026 | -0.023 | -0.022 | -0.020 | -0.015 | -0.009 | - | 0.013 | 0.020 | 0.024 | 0.028 | 0.030 | 0.032 | 0.035 | 0.036 | 0.049 | 0.074 | 0.077 | 0.079 | 0.099 | 0.125 | 0.141 |
| ROUND2_TKK | -0.038 | -0.036 | -0.035 | -0.033 | -0.027 | -0.022 | -0.013 | - | 0.008 | 0.011 | 0.015 | 0.017 | 0.019 | 0.023 | 0.024 | 0.037 | 0.061 | 0.065 | 0.066 | 0.086 | 0.113 | 0.129 |
| INRIA_Larlus | -0.046 | -0.044 | -0.042 | -0.040 | -0.035 | -0.029 | -0.020 | -0.008 | - | 0.004 | 0.007 | 0.009 | 0.011 | 0.015 | 0.016 | 0.029 | 0.053 | 0.057 | 0.059 | 0.079 | 0.105 | 0.121 |
| UVA_big5 | -0.050 | -0.047 | -0.046 | -0.044 | -0.039 | -0.033 | -0.024 | -0.011 | -0.004 | - | 0.004 | 0.006 | 0.008 | 0.011 | 0.013 | 0.025 | 0.050 | 0.053 | 0.055 | 0.075 | 0.101 | 0.118 |
| RWTH_GMM | -0.053 | -0.051 | -0.050 | -0.048 | -0.042 | -0.037 | -0.028 | -0.015 | -0.007 | -0.004 | - | 0.002 | 0.004 | 0.007 | 0.009 | 0.022 | 0.046 | 0.049 | 0.051 | 0.071 | 0.098 | 0.114 |
| TKK | -0.055 | -0.053 | -0.052 | -0.049 | -0.044 | -0.039 | -0.030 | -0.017 | -0.009 | -0.006 | -0.002 | - | 0.002 | 0.006 | 0.007 | 0.020 | 0.044 | 0.048 | 0.049 | 0.069 | 0.096 | 0.112 |
| RWTH_DiscHist | -0.057 | -0.055 | -0.054 | -0.052 | -0.046 | -0.041 | -0.032 | -0.019 | -0.011 | -0.008 | -0.004 | -0.002 | - | 0.003 | 0.005 | 0.018 | 0.042 | 0.046 | 0.047 | 0.067 | 0.094 | 0.110 |
| MUL_1v1 | -0.061 | -0.059 | -0.057 | -0.055 | -0.050 | -0.044 | -0.035 | -0.023 | -0.015 | -0.011 | -0.007 | -0.006 | -0.003 | - | 0.001 | 0.014 | 0.039 | 0.042 | 0.044 | 0.064 | 0.090 | 0.106 |
| RWTH_SparseHists | -0.062 | -0.060 | -0.059 | -0.056 | -0.051 | -0.046 | -0.036 | -0.024 | -0.016 | -0.013 | -0.009 | -0.007 | -0.005 | -0.001 | - | 0.013 | 0.037 | 0.041 | 0.043 | 0.062 | 0.089 | 0.105 |
| MUL_1vALL | -0.075 | -0.073 | -0.071 | -0.069 | -0.064 | -0.058 | -0.049 | -0.037 | -0.029 | -0.025 | -0.022 | -0.020 | -0.018 | -0.014 | -0.013 | - | 0.024 | 0.028 | 0.030 | 0.050 | 0.076 | 0.092 |
| AP06_Lee | -0.099 | -0.097 | -0.096 | -0.094 | -0.088 | -0.083 | -0.074 | -0.061 | -0.053 | -0.050 | -0.046 | -0.044 | -0.042 | -0.039 | -0.037 | -0.024 | - | 0.003 | 0.005 | 0.025 | 0.052 | 0.068 |
| INSARouen | -0.103 | -0.101 | -0.099 | -0.097 | -0.092 | -0.086 | -0.077 | -0.065 | -0.057 | -0.053 | -0.049 | -0.048 | -0.046 | -0.042 | -0.041 | -0.028 | -0.003 | - | 0.002 | 0.022 | 0.048 | 0.064 |
| UVA_weibull | -0.105 | -0.102 | -0.101 | -0.099 | -0.094 | -0.088 | -0.079 | -0.066 | -0.059 | -0.055 | -0.051 | -0.049 | -0.047 | -0.044 | -0.043 | -0.030 | -0.005 | -0.002 | - | 0.020 | 0.046 | 0.062 |
| Cambridge | -0.125 | -0.122 | -0.121 | -0.119 | -0.114 | -0.108 | -0.099 | -0.086 | -0.079 | -0.075 | -0.071 | -0.069 | -0.067 | -0.064 | -0.062 | -0.050 | -0.025 | -0.022 | -0.020 | - | 0.026 | 0.043 |
| Siena | -0.151 | -0.149 | -0.147 | -0.145 | -0.140 | -0.134 | -0.125 | -0.113 | -0.105 | -0.101 | -0.098 | -0.096 | -0.094 | -0.090 | -0.089 | -0.076 | -0.052 | -0.048 | -0.046 | -0.026 | - | 0.016 |
| AP06_Batra | -0.167 | -0.165 | -0.164 | -0.161 | -0.156 | -0.151 | -0.141 | -0.129 | -0.121 | -0.118 | -0.114 | -0.112 | -0.110 | -0.106 | -0.105 | -0.092 | -0.068 | -0.064 | -0.062 | -0.043 | -0.016 | - |

# Classification: Are errors “intuitive”?

- Class images:
  Highest ranked



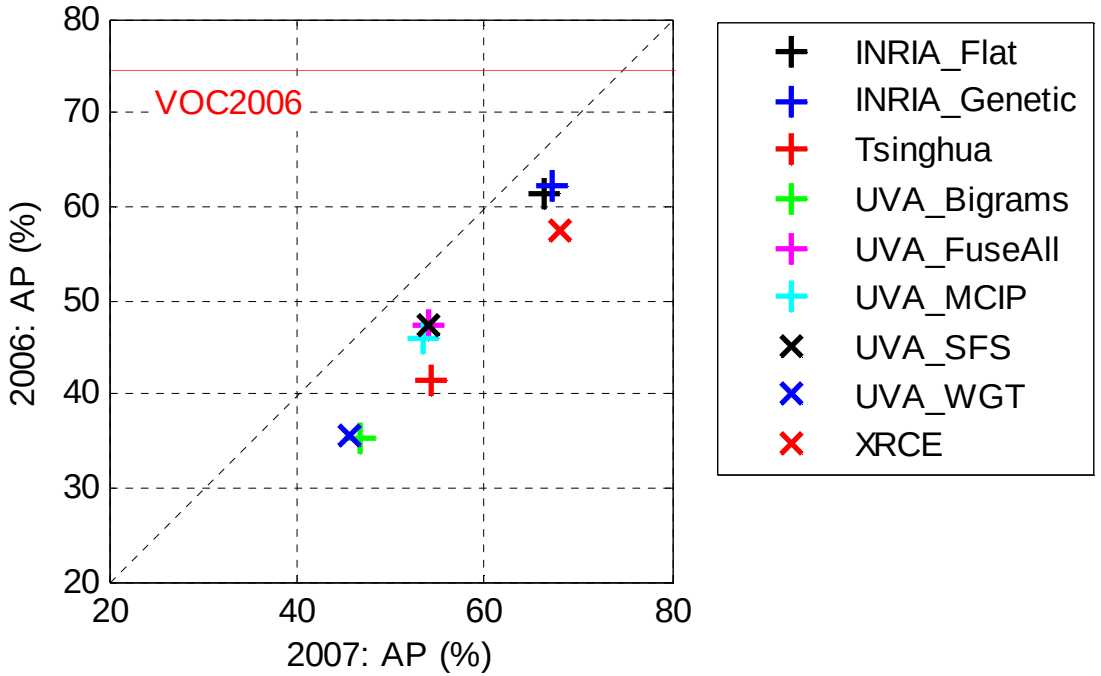
- Class images:
  Lowest ranked



- Non-class images:
  Highest ranked

- “Structured” Texture?

# Classification: Are methods getting better?

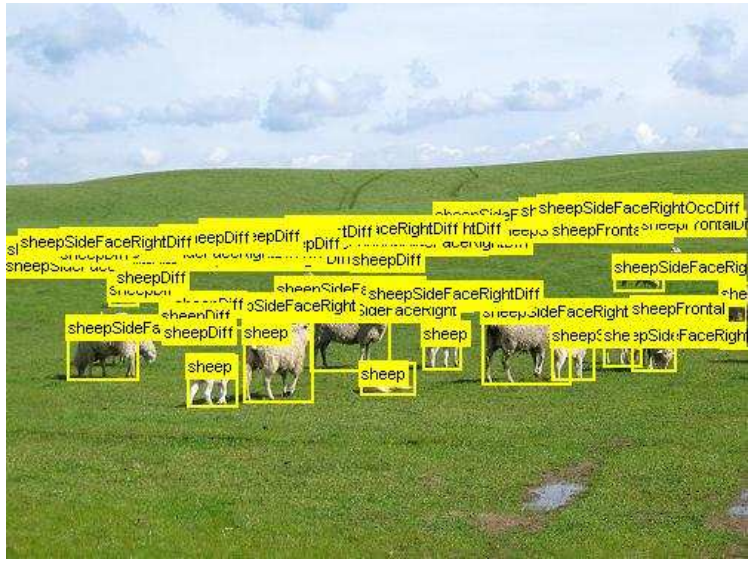



- High correlation between results on 2007 and 2006 test data
- Some evidence of "over-fitting" – no method equalled results when trained on 2006 data

# For Discussion...

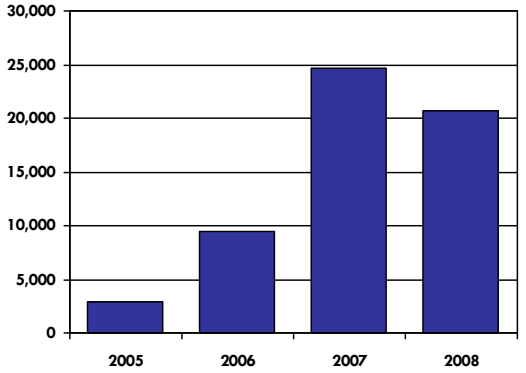
# Dataset

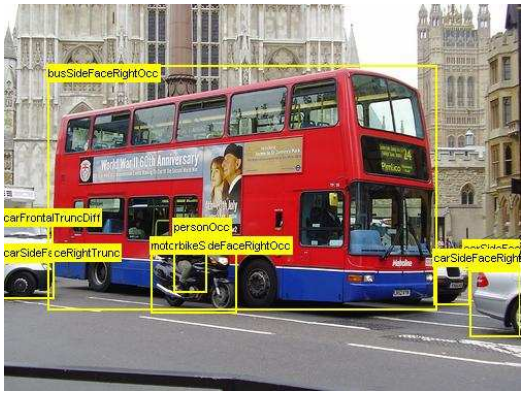
- Known Bias
  - Some bias due to keyword-based image collection
  - Images with only many small objects are discarded
  - Segmentation/pose data is biased towards simple scenes with larger objects
- Small Objects/Context
  - Objects unrecognizable in isolation are ignored in the evaluation but **are** included in the annotation



# Sustainability

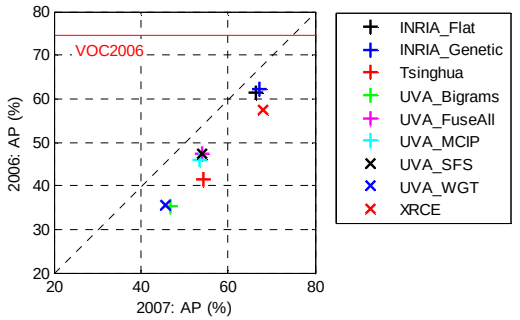
- Cost & Difficulty
  - Annotation is expensive: ~700 person hours for 2008
  - New (test) data is required each year to support withholding test annotation
  - Difficult to maintain high quality annotation with increased number of object classes ("cognitive load")
- Availability of Data
  - Becoming difficult to find examples of certain categories on flickr



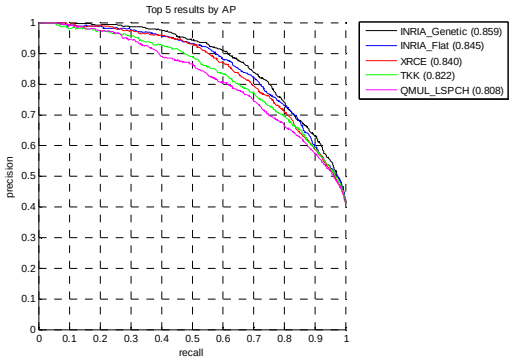



# Challenge

- "Longitudinal" Data
  - New test set every year makes measuring improvement difficult
  - Stop collecting more (test) data?




- "Pushing the curve"?
  - Are we encouraging incremental research?
  - 17 classification methods in 2007 were "bag of words"



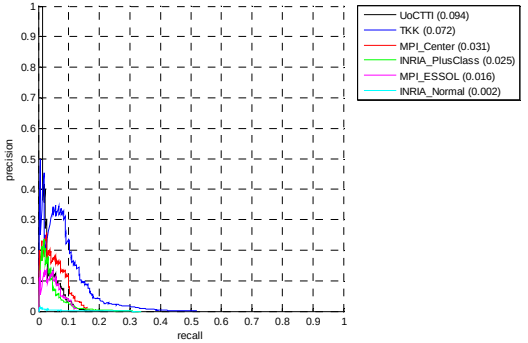
# Annotation

- Bounding Boxes?
  - More suitable for some objects than others...


- Alternatives?
  - Should we be annotating less data in more detail?
    - Polygons, “sketches”, parts, pixels, ...?
  - Should we be annotating more data in less detail?
    - Weak supervision e.g. keywords at image level?
  - Are we annotating the right data?
    - Video?

# Evaluation

- Useful to the community?
  - Are we measuring the right thing?
  - How to provide useful diagnostic information to guide research?
  - Is the data too difficult?




- "Taster" Challenges
  - Are the new challenges useful?
  - What other tasks should be introduced to stimulate research?