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

A new dataset is annotated annually. The annotation of the test set is withheld until after the challenge.

<table>
<thead>
<tr>
<th>Year</th>
<th>Images</th>
<th>Objects</th>
<th>Classes</th>
<th>Entries</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005</td>
<td>2,232</td>
<td>2,871</td>
<td>4</td>
<td>12</td>
<td>Collection of existing and some new data.</td>
</tr>
<tr>
<td>2006</td>
<td>5,304</td>
<td>9,507</td>
<td>10</td>
<td>25</td>
<td>Completely new dataset from flickr (+MSRC)</td>
</tr>
<tr>
<td>2007</td>
<td>9,963</td>
<td>24,640</td>
<td>20</td>
<td>28</td>
<td>Increased classes to 20. Introduced tasters.</td>
</tr>
<tr>
<td>2008</td>
<td>8,776</td>
<td>20,739</td>
<td>20</td>
<td></td>
<td>Added “occlusion” flag. Reuse of taster data. Release detailed results to support “meta-analysis”</td>
</tr>
</tbody>
</table>

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