Part II: VOC 2005-2012

The VOC years and legacy

Mark Everingham, Luc Van Gool
Chris Williams, John Winn
Andrew Zisserman
Yusuf Aytar, Ali Eslami

Outline

- Fine grained (easy and hard images) analysis
- Lessons on running challenges:
  - where we succeeded
  - where we could have done better

Per-image analysis

- Classification methods each assign a score to every image, and therefore induce a ranking on the images
- Can consider the ranks given to an image by the different methods
- Summarise ranks by their median value
- For true positives, show the images in the test dataset that
  1. belong to the class of interest, and
  2. are in the top 3 when ordered by the median rank given to them by the top methods

Top true positives (aeroplane)

- Top 50% of submissions evaluated of all years
- Evaluated on VOC2009 test data
- 396 images of aeroplanes out of 6650 test images in total in VOC2009

Low true positives (aeroplane)

- Top 50% of submissions evaluated of all years
- Evaluated on VOC2009 test data
- 396 images of aeroplanes out of 6650 test images in total in VOC2009

Top false positives (aeroplane)

- Top 50% of submissions evaluated of all years
- Evaluated on VOC2009 test data
- 396 images of aeroplanes out of 6650 test images in total in VOC2009

Top true positives (bicycle)

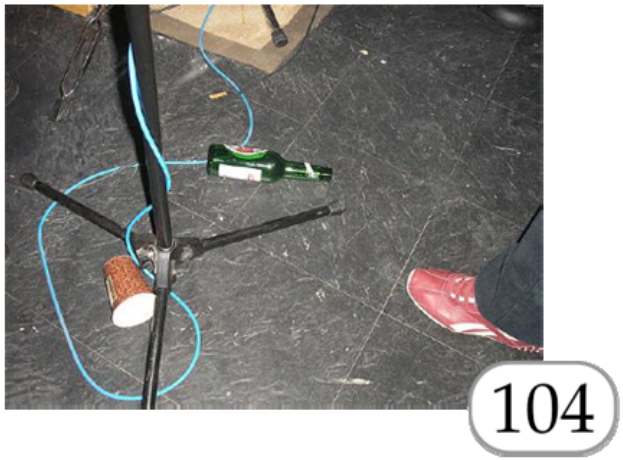
- Top 50% of submissions evaluated of all years
- Evaluated on VOC2009 test data
- 332 images of bicycles out of 6650 total in VOC2009

Low true positives (bicycle)

- Top 50% of submissions evaluated of all years
- Evaluated on VOC2009 test data
- 332 images of bicycles out of 6650 total in VOC2009

Top false positives (bicycle)

- Top 50% of submissions evaluated of all years
- Evaluated on VOC2009 test data
- 332 images of bicycles out of 6650 total in VOC2009

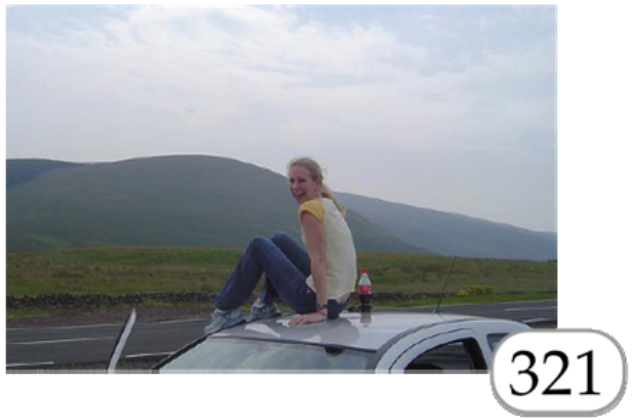


Top true positives (person)

- Top 50% of submissions evaluated of all years
- Evaluated on VOC2009 test data
- 2581 images of people out of 6650 total in VOC2009

Low true positives (person)

- Top 50% of submissions evaluated of all years
- Evaluated on VOC2009 test data
- 2581 images of people out of 6650 total in VOC2009

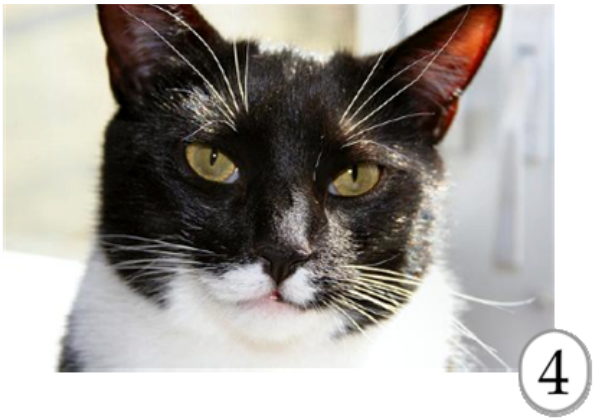


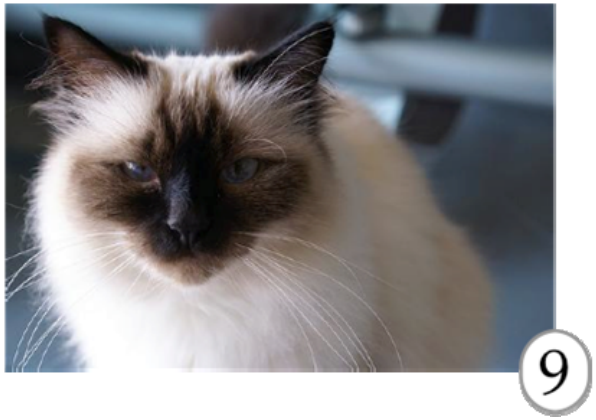
Top false positives (person)

- Top 50% of submissions evaluated of all years
- Evaluated on VOC2009 test data
- 2581 images of people out of 6650 total in VOC2009

Top true positives (cat)

- Top 50% of submissions evaluated of all years
- Evaluated on VOC2009 test data
- 538 images of cats out of 6650 total in VOC2009





Low true positives (cat)

- Top 50% of submissions evaluated of all years
- Evaluated on VOC2009 test data
- 538 images of cats out of 6650 total in VOC2009

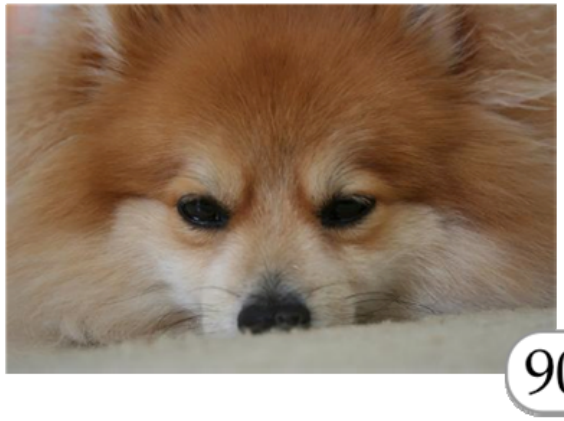
140


692


1001

Top false positives (cat)

- Top 50% of submissions evaluated of all years
- Evaluated on VOC2009 test data
- 538 images of cats out of 6650 total in VOC2009



Design choices: Things we got right …

1. Standard method of assessment

- Train/validation/test splits given
- Standard evaluation protocol – AP per class
- Software supplied
  - Includes baseline classifier/detector/segmenter
  - Runs from training to generation PR curve and AP on validation or test data out of the box
- Has meant that results on VOC can be consistently compared in publications

Design choices: Things we got right …

2. Evaluation of test data

Three possibilities:

1. Release test data and annotation (most liberal) and participants can assess performance
   - Cons: open to abuse

2. Release test data, but test annotation withheld - participants submit results and organizers assess performance (evaluation server)

3. No release of test data - participants have to submit software and organizers run this and assess performance
   - Cons: huge computational and software issues

Design choices: Things we got right …

3. Augmentation of dataset each year (up to 2011)

year	images	objects	Seg. objects
2008	4,340	10,363	2,369
2009	7,054	17,218	3,211
2010	10,103	23,374	4,203
2011	11,530	27,450	5,034
2012	11,530	27,450	6,929

- Has prevented over fitting on data
- 2008/9 datasets retained as subset of 2010-2012
  - Assignments to training/test sets maintained
  - So can measure progress from 2008 to 2012

Design choices: Things we got right …

4. The workshop

Recognized innovation as well as performance

Things we didn't get right: diversity

- Biggest risk of running any competition: **reduction in diversity of methods.**
  - New methods may be discarded before they mature, (because they don't beat the current mature methods)
  - Good strategy: do incremental improvements on last year's winning method
- Our solution:
  - Continually add new challenges
  - Individual challenges kept (largely) fixed, so we could track progress
  - BUT reduction in diversity on individual challenges

Boosting Diversity

- Another idea: use **boosting**
  - Attach weights to each test example
  - Increase weight of difficult test examples (that participants did poorly on in the previous year)
  - Compute weighted evaluation metrics
- This would:
  - Allow a challenge to be (essentially) fixed but still encourage diversity over time.
  - But: may focus attention on niche problems and lead to non-general solutions. Also: adds complexity.
  - Worth considering for future challenges?

Winston Churchill: “Democracy is the worst form of government except all the others that have been tried.”

Plans …

- Evaluation server to include banner/header results (to aid comparisons for reviewing etc) cf Middlebury

- Uses bootstrapping on rank test to determine equivalence class for methods

Successes

- Contributed to surge of interest in category recognition
- Contributed to establishing the importance of benchmarks (and efforts to refine them, e.g. Hoiem *et al*. ECCV 2012)
- PASCAL VOC mentioned in thousands of papers
- Have been able to measure steady performance increase in this area
- Felzenszwalb *et al*. DPM
- Combination of detection and classification

And, finally, thank you to the hundreds of participants that have taken part in the challenges over the years