OXFORD_VGG @ ILSVRC 2012

Karen Simonyan

Yusuf Aytar

Andrea Vedaldi

Andrew Zisserman

This is unpublished work.
Please cite this presentation or contact the authors
if you plan to make use of any of the ideas presented.

Our Approach

- Combine classification and detection in a cascade
  – class-specific bbox proposals
  – advanced features for proposal scoring
- Training in two stages:
  1. independent training
     - image classifiers
     - object detectors
  2. combination
     - object-level classifiers (bbox proposal scoring)
     - scores fusion

Our Approach

- Combine classification and detection in a cascade
  – class-specific bbox proposals
  – advanced features for proposal scoring
- Training in two stages:
  1. **independent training**
     - **image classifiers**
     - object detectors
  2. combination
     - object-level classifiers (bbox proposal scoring)
     - scores fusion

Image-Level Classification

Conventional approach: Fisher vector + linear SVM [1]

- Dense patch features
  – root-SIFT [2] & color statistics
  – augmentation with patch location (x,y) [3]
- Fisher vector (1024 Gaussians) => 135K-dim
- Compression using product quantization
- One-vs-rest linear SVM
  – early fusion: stacked root-SIFT FV and color FV (270K-dim)
  – Pegasos SGD

[1] Sanchez, Perronnin: "High-dimensional signature compression for large-scale image classification", CVPR 2011

[2] Arandjelovic, Zisserman: "Three things everyone should know to improve object retrieval ", CVPR 2012

[3] Sanchez et al.: "Modeling the Spatial Layout of Images Beyond Spatial Pyramids", PRL 2012

Classification: Comparison

Submission	Method	Error rate
SuperVision	DBN	0.16422
ISI	FV: SIFT, LBP, GIST, CSIFT	0.26172
XRCE/INRIA	FV: SIFT and colour 1M-dim features	0.27058
OXFORD_VGG	FV: SIFT and colour 270K-dim features **(classification only, no fusion)**	0.27302

9.8%

1.1%

- Saturation of FV-based approaches
- Adding more off-the-shelf features or increasing dimensionality does not help much

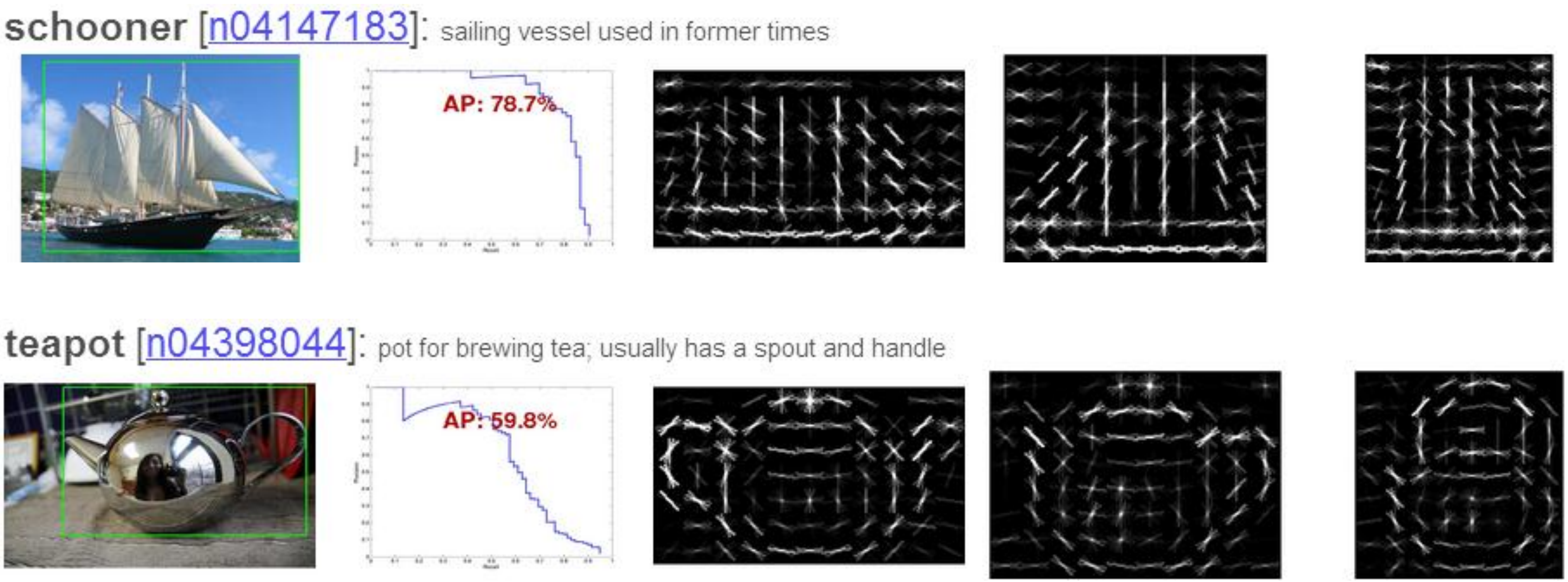
Our Approach

- Combine classification and detection in a cascade
  - class-specific bbox proposals
  - advanced features for proposal scoring
- Training in two stages:
  1. independent training
     - image classifiers
     - **object detectors**
  2. combination
     - object-level classifiers (bbox proposal scoring)
     - scores fusion

Detection: DPMs

Discriminatively trained part based models [1]

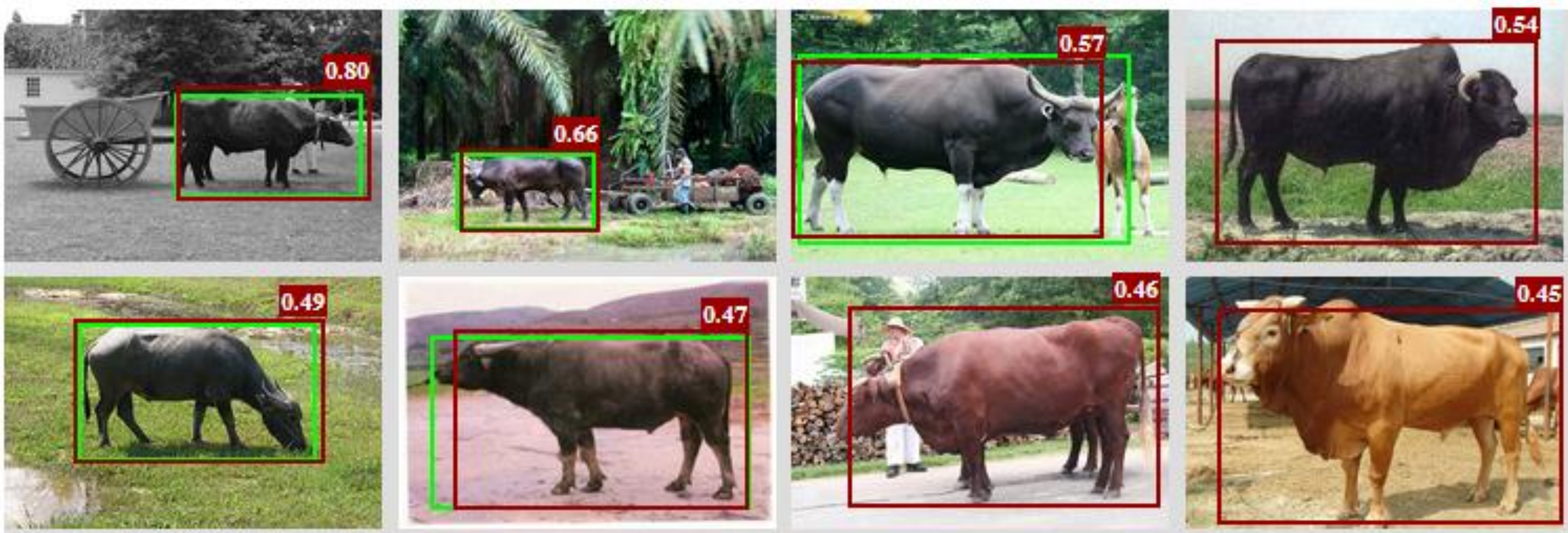
- 3 components (aspects)
- no parts (root filters only)



[1] Felzenszwalb et al.: "Object Detection with Discriminatively Trained Part Based Models", PAMI 2010

Semi-Supervised Learning

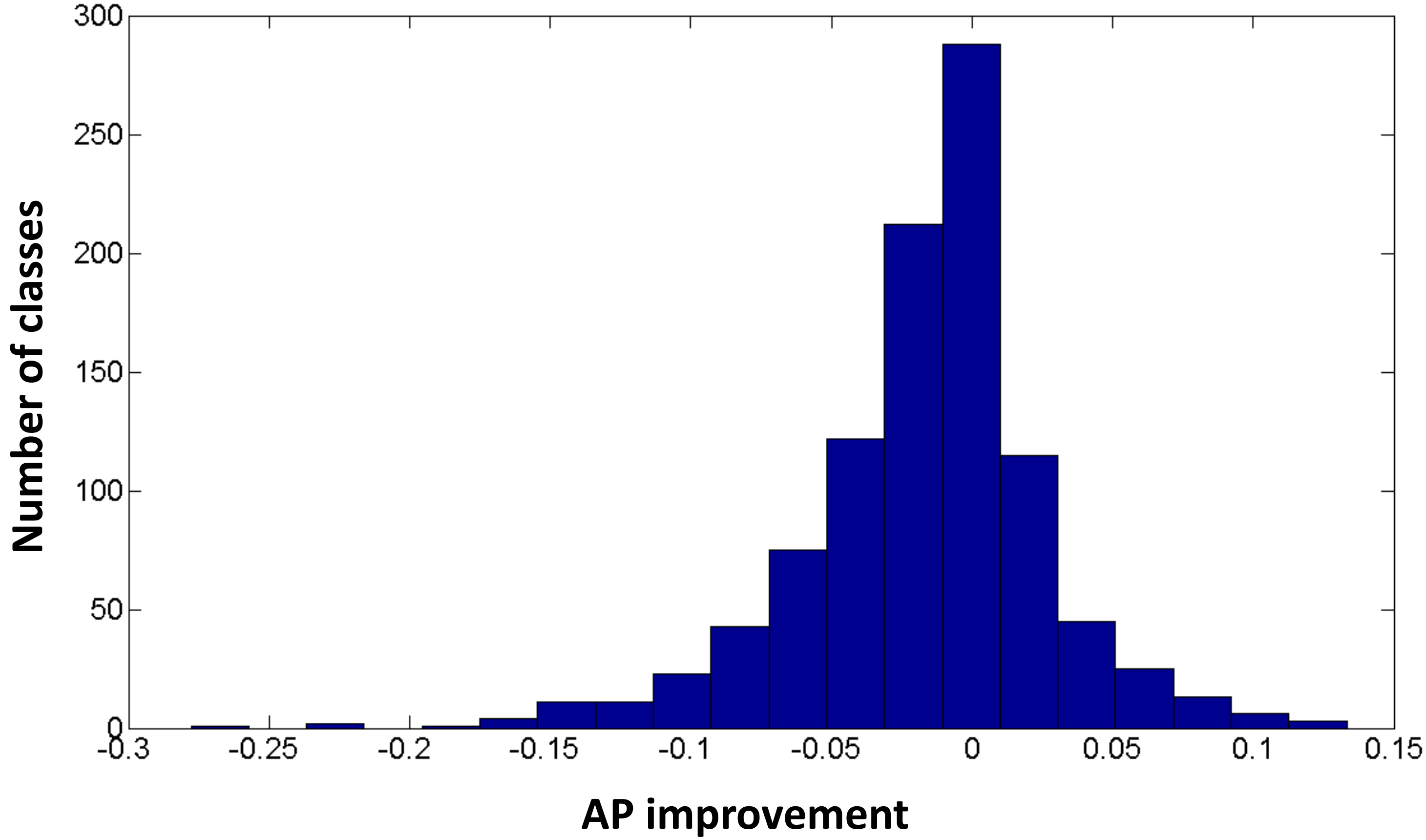
- Ground-truth bboxes available for only ~42% training images
- Training set augmentation:
  1. train detectors on ground-truth bboxes
  2. get more positives by detection on the rest of the training set



top-scored training set detections:
red – detected bbox; green – ground-truth bbox (if available)

SSL: Performance Improvements

- for 329 classes AP is improved (+2.4% on average)
- for the rest of the classes – training on ground-truth only

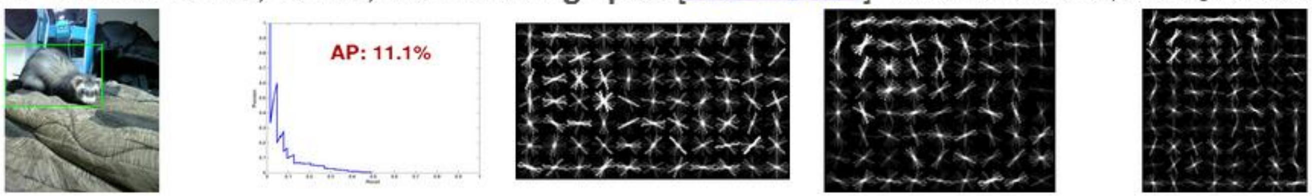


Quality of DPMs

Evaluation on the validation set

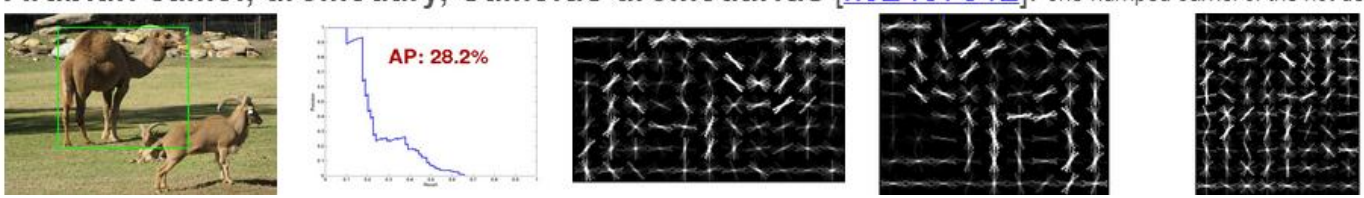
- AP in [0; 25%): 582 detectors

black-footed ferret, ferret, Mustela nigripes [n02443484]: musteline mammal of prairie regions of United



- AP in [25%; 50%]: 338 detectors

Arabian camel, dromedary, Camelus dromedarius [n02437312]: one-humped camel of the hot deserts

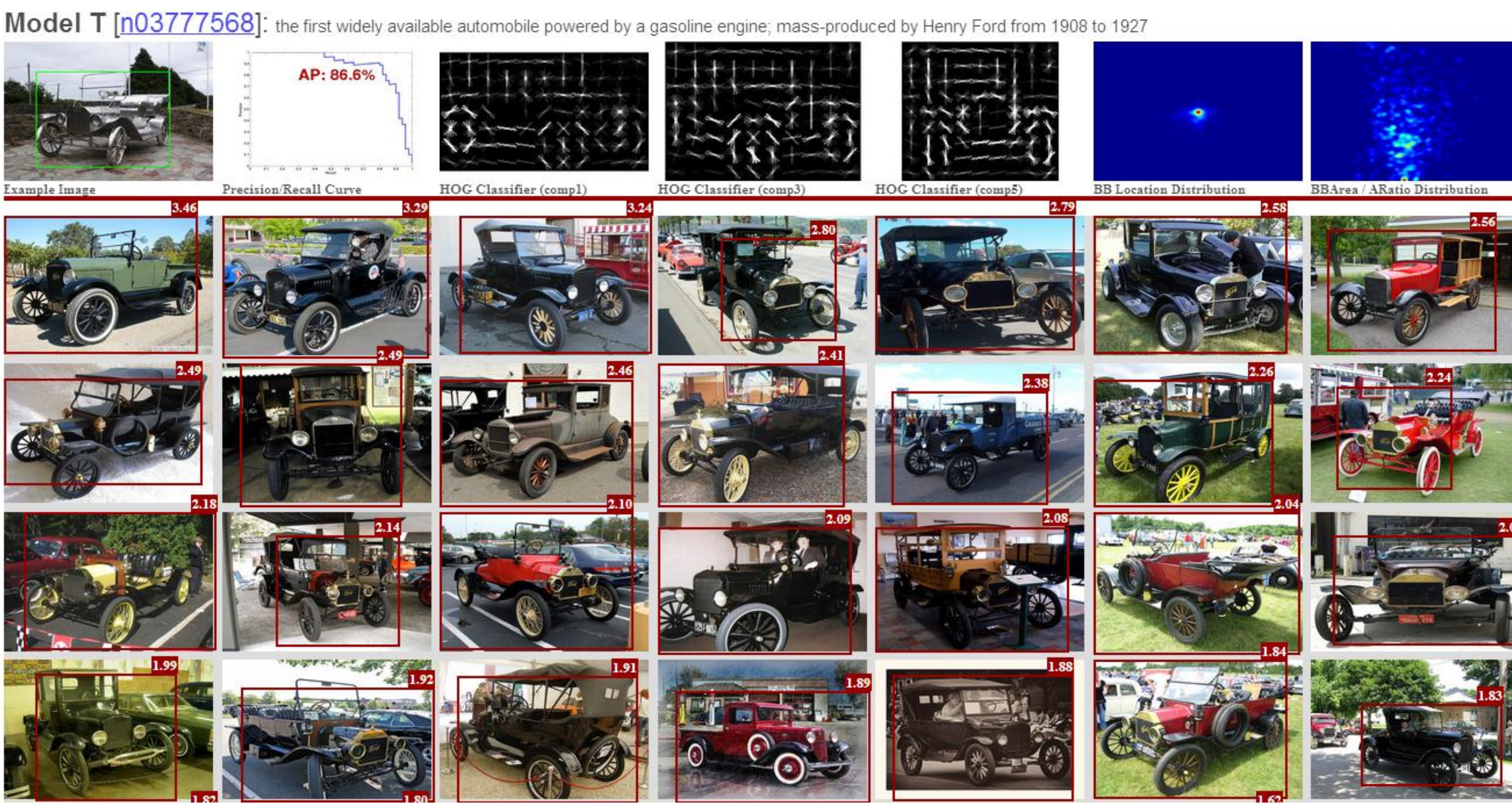


- AP in (50%; 100%]: 80 detectors

Leonberg [n02111129]: a large dog (usually with a golden coat) produced by crossing a St Bernard and a Newfoundland

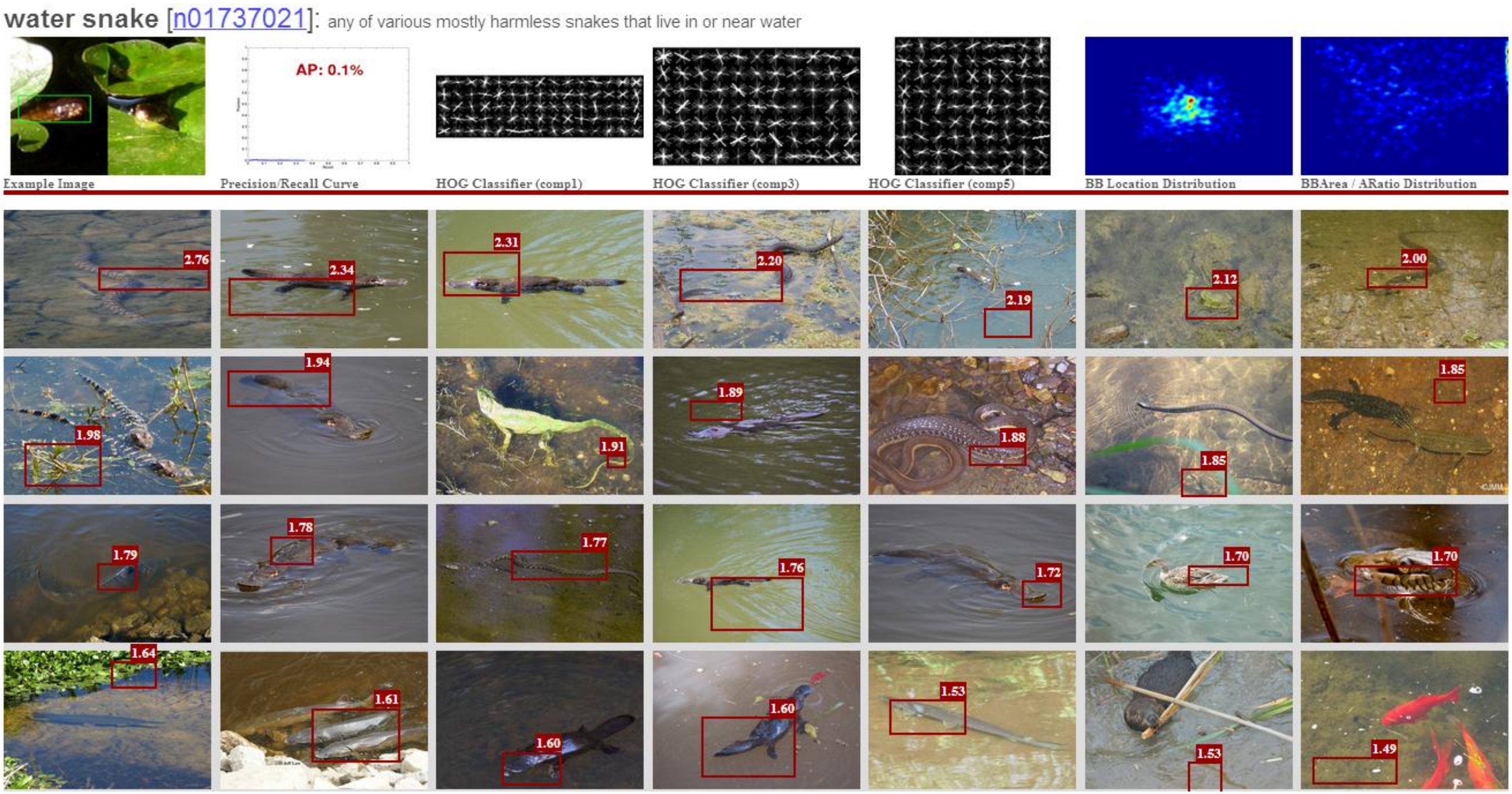
Best Detector (86.6% AP)

Strongly defined, unique shape



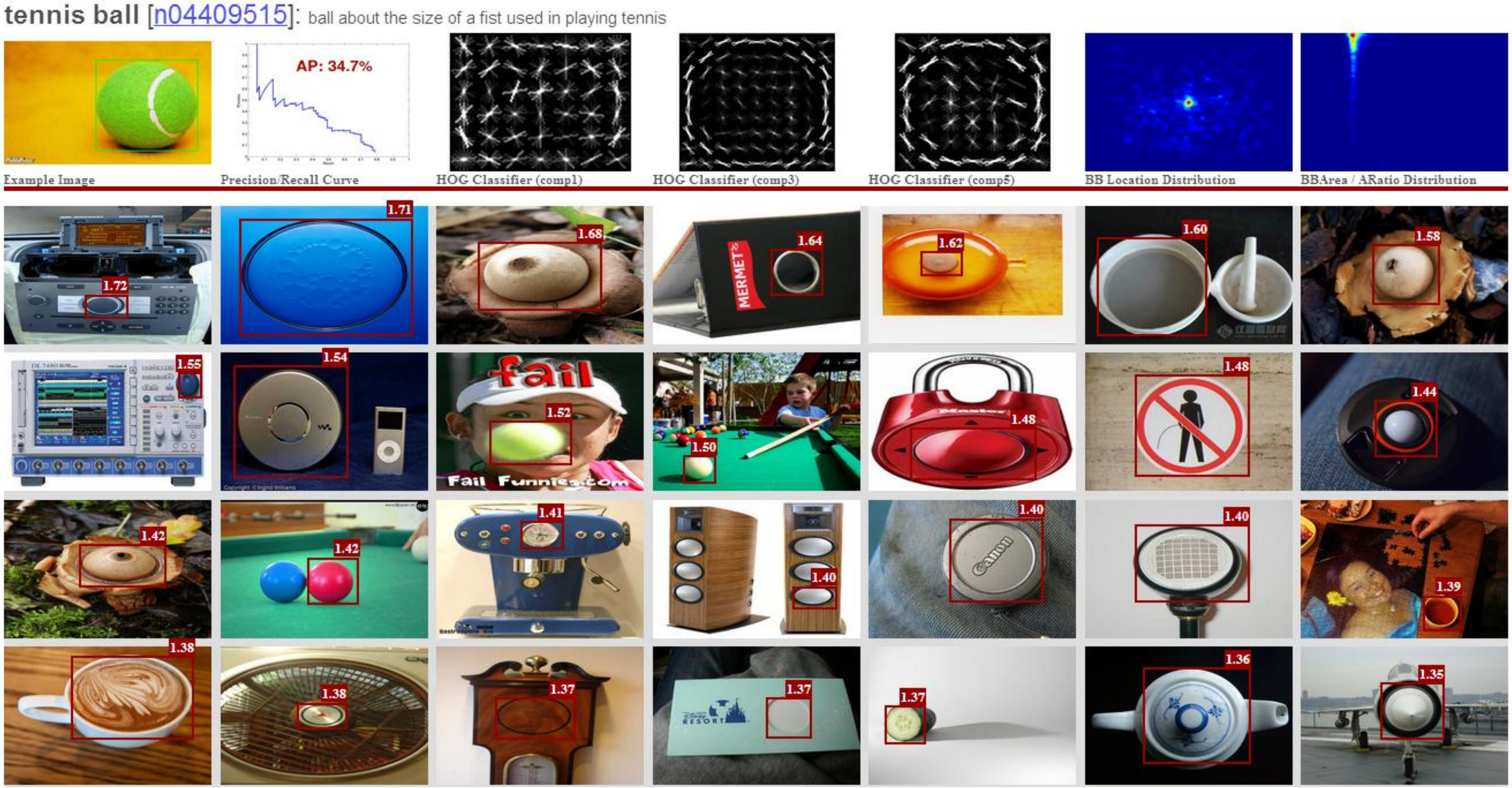
DPM Problems

- HOG models are not appropriate for certain classes
  - large variability in shape (e.g. reptiles)



DPM Problems

- Ambiguity between structurally similar classes
  - similar shape, but different appearance (e.g. fruit, dog breeds)

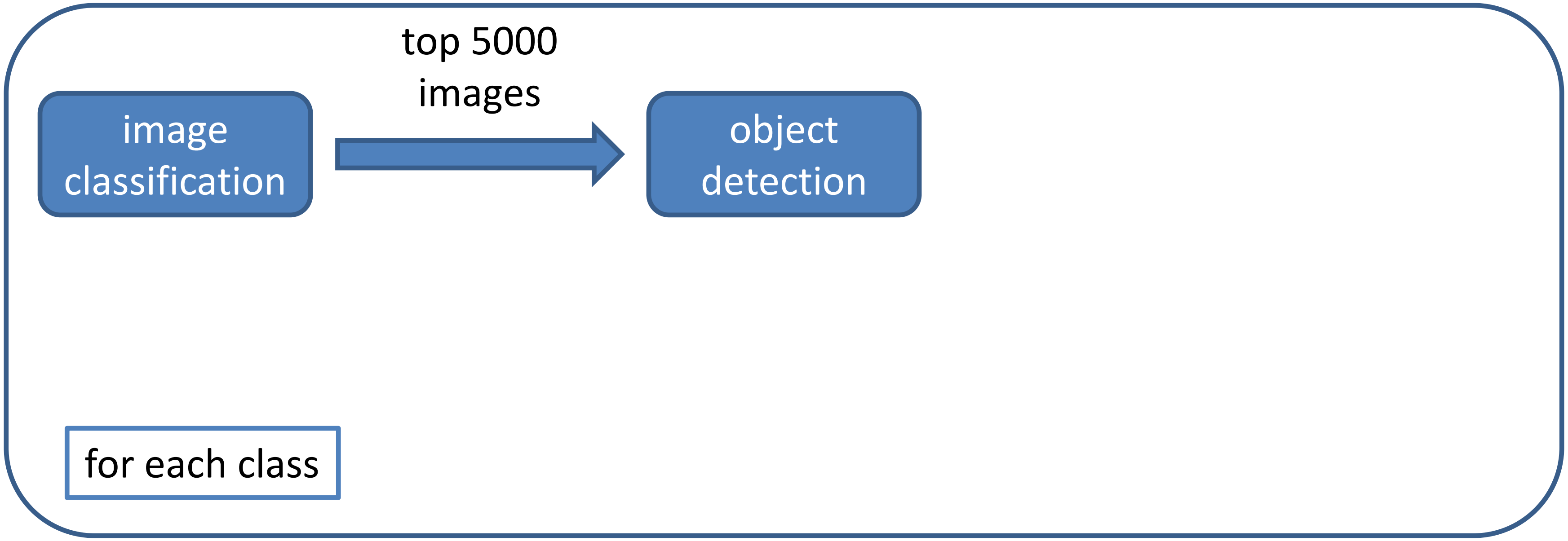


Our Approach

- Combine classification and detection in a cascade
  - class-specific bbox proposals
  - advanced features for proposal scoring
- Training in two stages:
  1. independent training
     - image classifiers
     - object detectors
  2. **combination**
     - **object-level classifiers (bbox proposal scoring)**
     - **scores fusion**

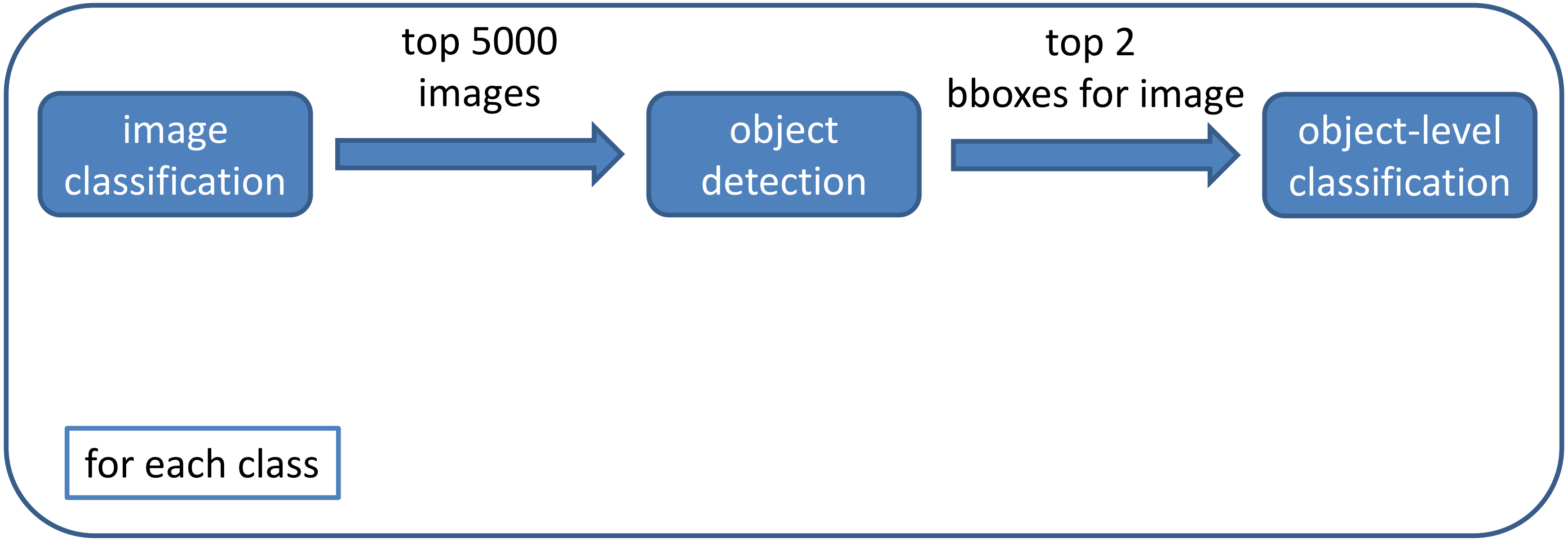
Applying DPM at Large Scale

- DPMs can provide good bbox predictions, but too slow
  - 1K classes x 100K test images = 100M sliding window runs
- Use classification to drive detection → speed-up
  - classification recall is quite high (90.7% at top 5%)
  - object detection on top 5000 (5%) images of each class



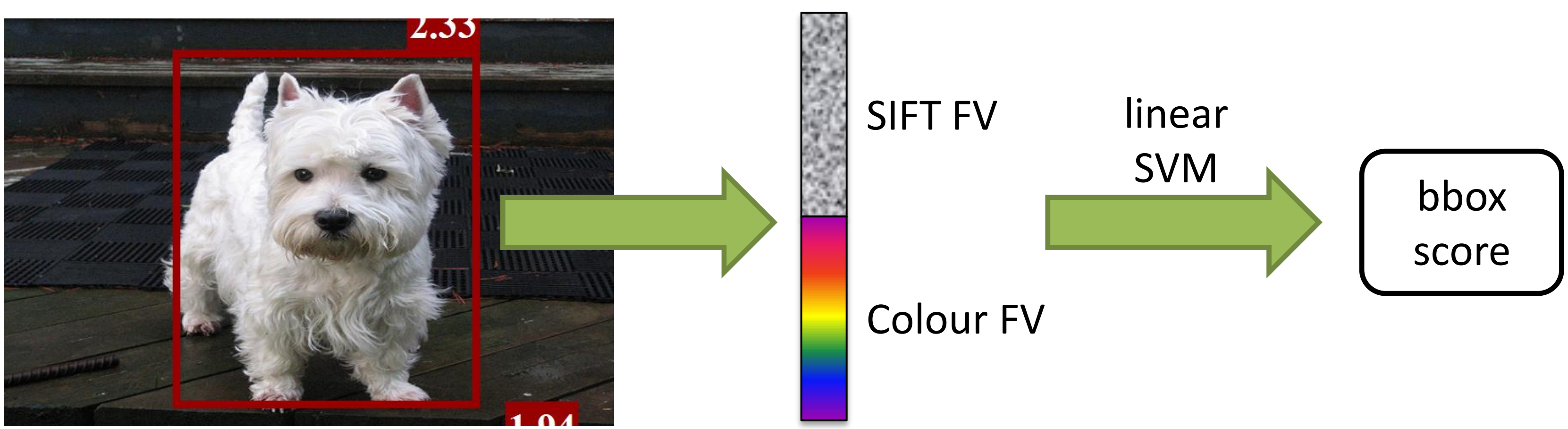
Bounding Box Proposals

- Top DPM detections are used as proposals
  - top 2 bboxes used in this submission
- Proposals are scored using more complex models
  - affordable for a few boxes



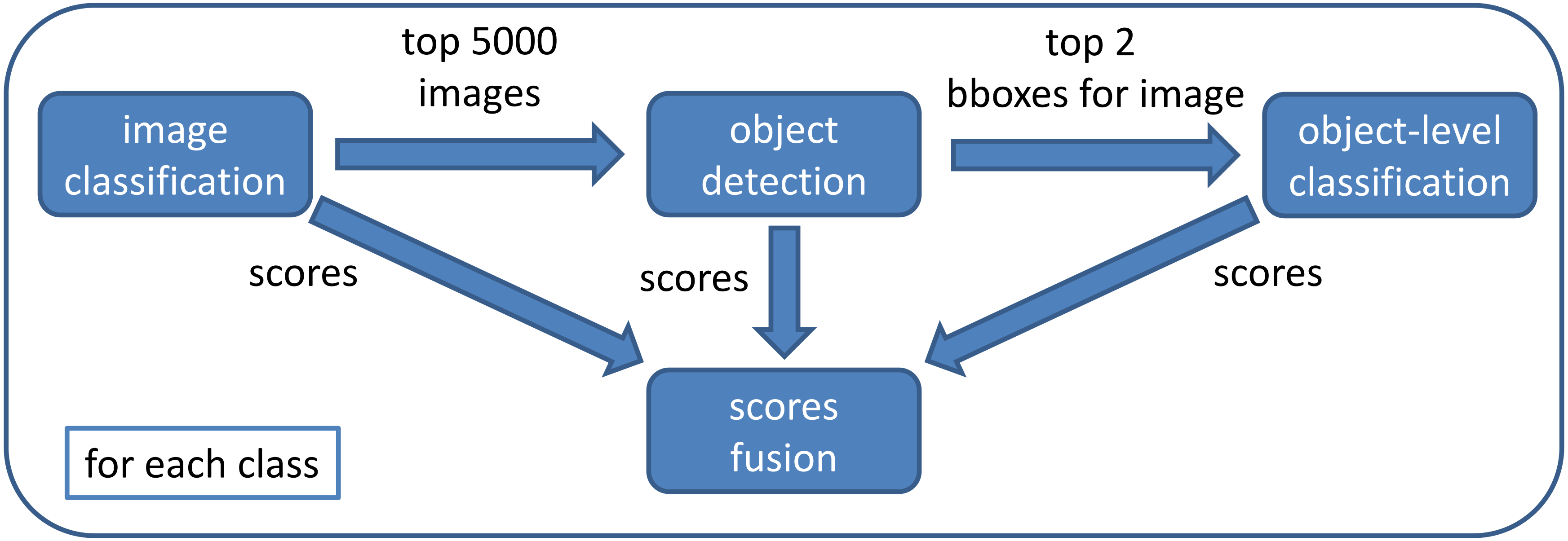
Object-Level Classification

- High-dimensional model
  - linear SVM with features as in image classification: DSIFT-FV & Color-FV (270K-dim.)
  - accounts for bbox-level texture & color cues
- Training set
  - training set positives
  - ⅓ of validation negatives (top 2 bboxes for each image)



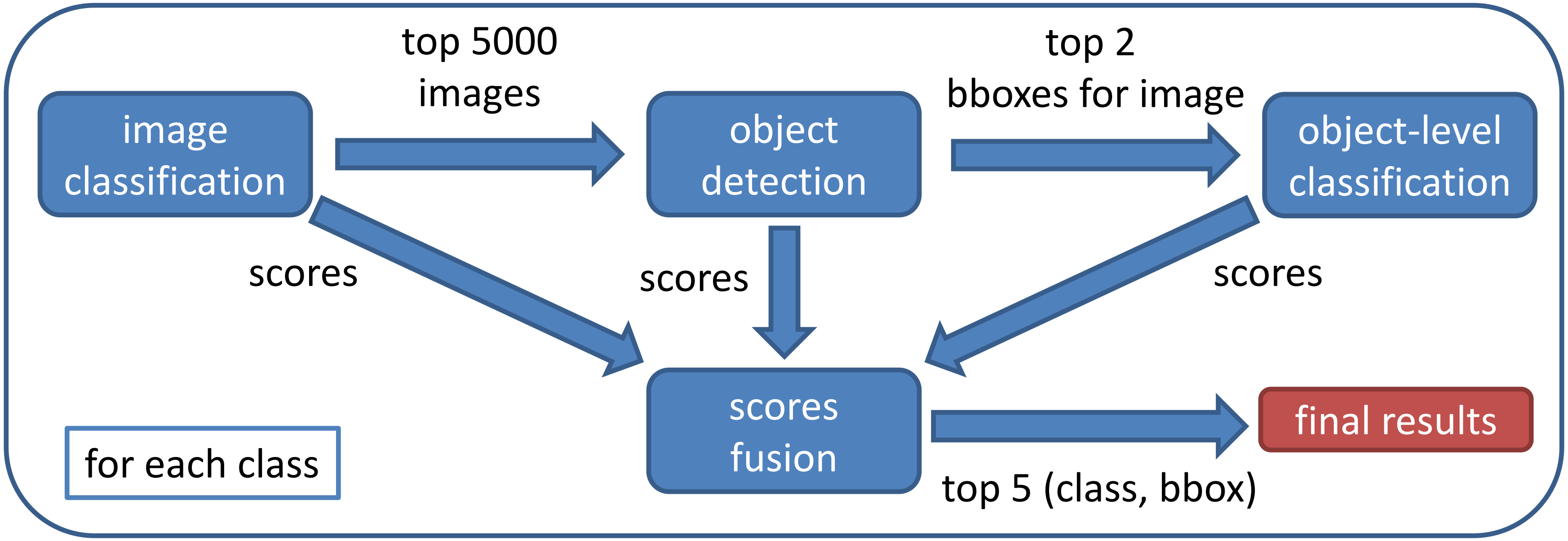
Scores Fusion

- Three scores are fused into a single one
  - fused score corresponds to object class and bbox



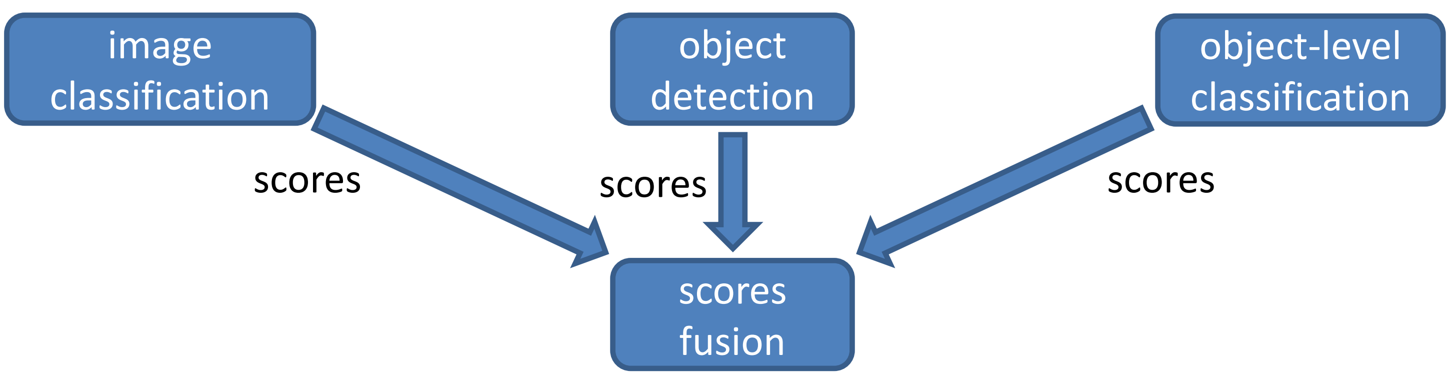
Scores Fusion

- Three scores are fused into a single one
  - fused score corresponds to object class and bbox
- Top 5 classes with bboxes determined by ranking on the (calibrated) fused scores
  - each image is in top 5000 of ≥ 10 classes, so top 5 is feasible



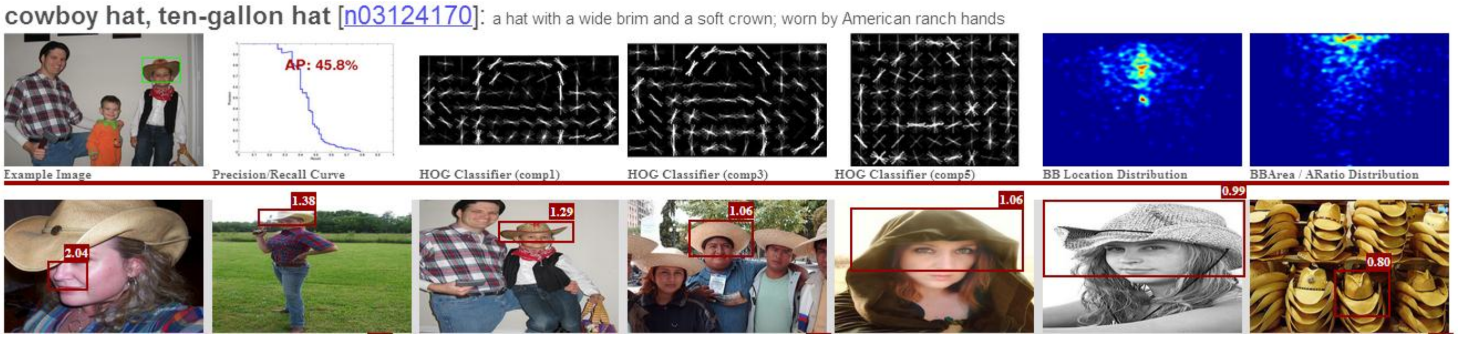
Scores Fusion: Learning

- Three complementary cues:
  - image-level classification score (dense SIFT & color)
  - object-level DPM score (HOG local shape information)
  - object-level classification score (dense SIFT & color)
- Fusion using linear combination of 3 scores
  - weights trained on the validation set using linear SVM



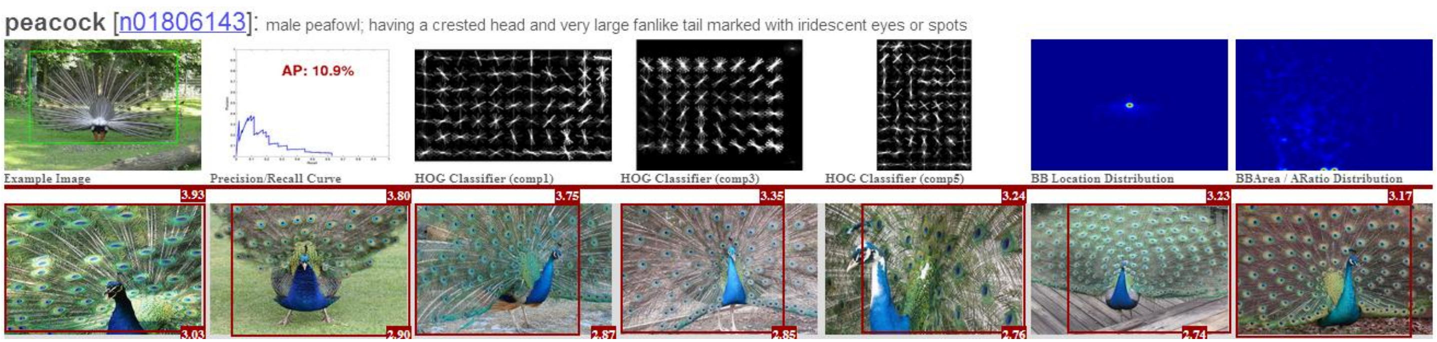
Is Fusion Helpful for Classification?

- It helps if objects occupy a small area and can be detected well



+25% AP

- It doesn't help if objects occupy the whole image
  - we use the same features



Is Fusion Helpful for Detection?

- What confuses DPM can be less ambiguous for fine-level classification

left: best bbox according to DPM; right: best bbox after scores fusion

Classification: Comparison

Submission	Method	Error rate
SuperVision	DBN	0.16422
ISI	FV: SIFT, LBP, GIST, CSIFT	0.26172
OXFORD_VGG	fusion of classification & detection	0.26979
XRCE/INRIA	FV: SIFT and colour 1M-dim features	0.27058
OXFORD_VGG	classification only FV: SIFT and colour 270K-dim features	0.27302

0.3%

- Slight improvement in classification accuracy
- Classification is already doing well for its class of methods

Detection: Comparison

Submission	Method	Error rate
SuperVision	DBN	0.341905
OXFORD_VGG	fusion of classification & detection, 2 DPM bbox proposals	0.500342
OXFORD_VGG	fusion of classification & detection, 1 DPM bbox proposal	0.522189
OXFORD_VGG	baseline: detection of top-5 classes based on classification	0.529482

2.9%

- Fusion brings a noticeable improvement compared to the baseline
- Using more proposals (2 vs 1) gives better results

Proposal Generation Approaches

- Class-dependent bbox proposals
  - 2 proposals for (class, image) → ~100 proposals/image
  - requires training
  - quality depends on the learned model
- Class-independent bbox proposals, e.g. "selective search" [1]
  - higher number of proposals (~1500 proposals/image)
  - makes very generic assumptions of object appearance
    - colour/texture uniformity
- Might complement each other

[1] van de Sande et al.: "Segmentation As Selective Search for Object Recognition", ICCV 2011

Summary

- Our framework allows for
  - high-quality class-specific bbox proposals (using DPM)
    - works well for classes with well-defined shapes
  - computationally complex features (FV) for bbox scoring
    - combination of various visual cues
- Future work
  - improve detection for classes with weakly-defined shapes
  - better low-level features