This is unpublished work.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
  – 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

## Classification: Comparison

<table>
<thead>
<tr>
<th>Submission</th>
<th>Method</th>
<th>Error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>SuperVision</td>
<td>DBN</td>
<td>0.16422</td>
</tr>
<tr>
<td>ISI</td>
<td>FV: SIFT, LBP, GIST, CSIFT</td>
<td>0.26172</td>
</tr>
<tr>
<td>XRCE/INRIA</td>
<td>FV: SIFT and colour 1M-dim features</td>
<td>0.27058</td>
</tr>
<tr>
<td>OXFORD_VGG</td>
<td>FV: SIFT and colour 270K-dim features</td>
<td>0.27302</td>
</tr>
</tbody>
</table>

- 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)

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

Model T [n03777568]: the first widely available automobile powered by a gasoline engine, mass-produced by Henry Ford from 1908 to 1927.
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 $\rightarrow$ speed-up
  – classification recall is quite high (90.7% at top 5%)
  – object detection on top 5000 (5%) images of each class

For each class:
- Top 5000 images
- Image classification
- Object detection
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

Diagram:

- Image classification
- Top 5000 images
- Object detection
- Top 2 bboxes for image
- Object-level classification

for each class
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
  – $\frac{1}{3}$ 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 $\geq 10$ 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
  cowboy hat, ten-gallon hat [n03124170]: a hat with a wide brim and a soft crown, worn by American ranch hands
  +25% AP

• It doesn't help if objects occupy the whole image
  – we use the same features
  peacock [n01806143]: male peafowl, having a crested head and very large fanlike tail marked with iridescent eyes or spots
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

<table>
<thead>
<tr>
<th>Submission</th>
<th>Method</th>
<th>Error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>SuperVision</td>
<td>DBN</td>
<td>0.16422</td>
</tr>
<tr>
<td>ISI</td>
<td>FV: SIFT, LBP, GIST, CSIFT</td>
<td>0.26172</td>
</tr>
<tr>
<td>OXFORD_VGG</td>
<td>fusion of classification &amp; detection</td>
<td>0.26979</td>
</tr>
<tr>
<td>XRCE/INRIA</td>
<td>FV: SIFT and colour 1M-dim features</td>
<td>0.27058</td>
</tr>
<tr>
<td>OXFORD_VGG</td>
<td>classification only FV: SIFT and colour 270K-dim features</td>
<td>0.27302</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Submission</th>
<th>Method</th>
<th>Error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>SuperVision</td>
<td>DBN</td>
<td>0.341905</td>
</tr>
<tr>
<td>OXFORD_VGG</td>
<td>fusion of classification &amp; detection, 2 DPM bbox proposals</td>
<td>0.500342</td>
</tr>
<tr>
<td>OXFORD_VGG</td>
<td>fusion of classification &amp; detection, 1 DPM bbox proposal</td>
<td>0.522189</td>
</tr>
<tr>
<td>OXFORD_VGG</td>
<td>baseline: detection of top-5 classes based on classification</td>
<td>0.529482</td>
</tr>
</tbody>
</table>

- 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

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