Object category localization with incomplete supervision

Jakob Verbeek

LEAR team, INRIA, Grenoble, France

Joint work with: Gokberk Cinbis and Cordelia Schmid
Object category localization

- Locate object category instances by means of bounding box

- Supervised learning setup:
  - Training images with bounding-box annotations of object instances
  - Learn a binary classifier: windows are a category instance or not

- Numerous applications
  - Surveillance
  - Traffic safety: autonomous or assisted driving systems
  - ...
Why learning from incomplete supervision?

- Bounding boxes are much more expensive to get than image labels
- Weakly supervised learning only uses image-wide labels
  - For positive images we only know there's at least one instance, but we don't know how many and where they are
  - Less detail in supervision than in target outputs
Presentation outline

- Preliminaries on object localization
  - Challenges
  - Representations
  - Search and learning

- Learning with incomplete supervision
  - Multiple instance learning approach
  - Multi-fold training to improve performance
  - Object instance hypothesis refinement

- Experimental evaluation results
Challenging factors in object detection

- Intra-class appearance variation
  - Objects deformation due to pose
  - Transparency: e.g. bottles
  - Sub-categories: e.g. ferry vs yacht

- Scene composition
  - Heavy occlusions: e.g. tables and chairs
  - Clutter: coincidental image content present in bounding box

- Imaging conditions
  - viewpoint, scale, lighting conditions
State-of-the-art visual representations (1/2)

- Fischer vector image representation [Sanchez et al., IJCV, 2013]
  - Represent data with gradient of log-likelihood of generative model
- Densely sampled SIFT descriptors modeled with Gaussian mixture
- Encode an image by gradient w.r.t. means and variances: 2KD vector
  - Results in a 140K dimensional signature

\[
\nabla_{\mu_k} \ln p(x_{1:N}) = \frac{1}{\sigma_k \sqrt{\pi_k}} \sum_{n=1}^{N} p(k|x_n)(x_n - \mu_k)
\]
\[
\nabla_{\sigma_k} \ln p(x_{1:N}) = \frac{1}{\sigma_k \sqrt{2 \pi_k}} \sum_{n=1}^{N} p(k|x_n)((x_n - \mu_k)^2 - \sigma_k^2)
\]
State-of-the-art visual representations (2/2)

- Use a Convolutional Neural Network as a feature extraction method for object detection [R-CNN, Girschik et al., 2013]
- Trained on 1 million images of 1000 categories (ImageNet 2012)
- Caffe framework [Jia et al., caffe.berkeleyvision.org]
- Use last shared layer as a 4K dimensional representation
How to avoid exhaustive sliding window search

- **Branch-and-bound techniques**

  - Sliding window
    - (Viola and Jones 2002; Felzenszwalb et al. 2008, ...)
  - Branch & bound
    - (Lampert et al. 2008; Lehmann et al. 2013)
  - Selective Search
    - (Alexe et al. 2010; Sande et al. 2011)
Search: restricted scanning of bounding box space

- Selective search method [Uijlings et al., IJCV, 2013]
  - Unsupervised multi-resolution hierarchical segmentation
  - Detections proposals generated as bounding box of segments
  - 1500 windows per image suffice to cover over 95% of true objects with sufficient accuracy
Presentation outline

- Preliminaries on object localization
  - Challenges
  - Representations
  - Search and learning

- Learning with incomplete supervision
  - Multiple instance learning approach
  - Multi-fold training to improve performance
  - Object instance hypothesis refinement

- Experimental evaluation results
Learning from incomplete supervision

- A joint identification problem:
  - Locating object instances in positive images
  - Learning detector from positive and negative examples
State-of-the-art weakly-supervised detector training

- Vast majority of work relies on multiple-instance learning

- Approaches vary in terms of
  - Initialization strategy
  - Object descriptors and detector
  - Utilization of pair-wise window similarities

- Some alternative recent approaches are based on topic models, e.g. LDA
The multiple instance learning (MIL) approach

- Examples come in labeled “bags”
  - Positive bags contain at least one positive sample
  - Negative bags only contain negative samples

- Multiple Instance SVM
  - Initialize initial selection of samples from positive bags
  - Train SVM with selection
  - Select top scoring sample in each positive bag
  - Repeat until convergence

[Dietterich et al., Artif. Intell., 1997]
[Andrews et al., NIPS 2002]
Multiple instance learning in practice...

- Converges rapidly to poor local optima
The problems in multiple instance learning

- Given a trained detector, consider score of windows that
  - do not match true objects
  - do match true objects
  - were used as positive samples to train the detector (might be wrong)
Problems in standard multiple instance learning

- Our window descriptors are high dimensional
  - Descriptors are L2 normalized
  - Most pairs are near orthogonal, i.e. near-zero dot products

- Linear classifier score is weighted sum of dot products
  \[ w^T x = \sum_i \alpha_i (x_i^T x) \]

- Classifier scores much higher for positive windows used in training
  - This causes the degenerate re-localization behavior
Multi-fold training for multiple instance learning

- Separate sets of positive images for training and re-localization
  - Negative images do not need to be split, since no relocalization there

- Repeat two steps
  - Partition positive training images into K folds
  - For fold $k = 1, \ldots, K$
    - Train detector from all training images, except those in fold $k$
    - Select top-scoring window in each positive image in fold $k$

- Avoids the re-localization bias since images used for training and re-localization are always different
Multi-fold training for multiple instance learning algorithm

**Algorithm 1** — Multi-fold weakly supervised training

1) Initialization: positive and negative examples are set to entire images
2) For iteration $t = 1$ to $T$
   a) Divide positive images randomly into $K$ folds.
   b) For $k = 1$ to $K$
      i) Train using positive examples in all folds but $k$, and all negative examples.
      ii) Re-localize positives by selecting the top scoring window in each image of fold $k$ using this detector.
   c) Train detector using re-localized positives and all negative examples.
   d) Add new negative windows by hard-negative mining.
3) Return final detector and object windows in train data.
Multi-fold training for multiple instance learning

- Resolves the degenerate re-localization of standard MIL training
Limitation of weakly supervised learning

- Weakly supervised learning learns the most discriminative pattern between the positive and negative images.
- These patterns may correspond to parts instead of full objects.
  - For example, the faces of cats and dogs due to body poses.
Hypothesis refinement using low-level contours

- Encourage object hypotheses to align with long image contours
  - Using efficient contour alignment score [Zitnick & Dollar, ECCV'14]

- After multi-fold training iterations: use weighted combination of detection and contour alignment score

- Final detector trained using the refined hypotheses
Presentation outline

- Preliminaries on object localization
  - Challenges
  - Representations
  - Search and learning

- Learning with incomplete supervision
  - Multiple instance learning approach
  - Multi-fold training to improve performance
  - Object instance hypothesis refinement

- Experimental evaluation results
Evaluations based on PASCAL VOC'07 benchmark

- Most challenging dataset for weakly supervised detection
Evaluation protocols

- Localization success on positive training images
  - Fraction of images with correct localization (CorLoc) [Deselaers et al., PAMI 2012]
- Standard PASCAL-VOC detection Average Precision (AP) on test set
- Both measures averaged over 20 different object categories

- Detection declared a success if highly overlapping with ground-truth
  - Intersection-over-union of window areas larger than 50%

IoU = 0.5
IoU = 0.7
IoU = 0.9
Evaluation of multi-fold training

- Comparison of standard MIL training and multi-fold strategy

- Multi-fold training improves both performance measures using either Fisher vector or CNN features

<table>
<thead>
<tr>
<th></th>
<th>Standard</th>
<th>Multi-fold</th>
</tr>
</thead>
<tbody>
<tr>
<td>CorLoc</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FV</td>
<td>29.7</td>
<td>38.8 (+9.1)</td>
</tr>
<tr>
<td>CNN</td>
<td>41.2</td>
<td>45.0 (+3.8)</td>
</tr>
<tr>
<td>Detection AP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FV</td>
<td>15.5</td>
<td>22.4 (+6.9)</td>
</tr>
<tr>
<td>CNN</td>
<td>24.3</td>
<td>25.9 (+1.6)</td>
</tr>
</tbody>
</table>
Evaluation of multi-fold training

- CorLoc over the re-training / re-localization iterations
- Iteration n: n-th iteration after initialization from full image
Window refinement and combining features

- Contour alignment score improves performance
- Combining features boosts performance

<table>
<thead>
<tr>
<th>Refinement</th>
<th>No</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Detection AP</td>
<td></td>
</tr>
<tr>
<td>FV</td>
<td>22.4</td>
<td>23.3 (+0.9)</td>
</tr>
<tr>
<td>CNN</td>
<td>25.9</td>
<td>28.6 (+2.7)</td>
</tr>
<tr>
<td>FV+CNN</td>
<td>27.4</td>
<td>30.2 (+2.8)</td>
</tr>
</tbody>
</table>

- Classes with largest improvements due to contour alignment

<table>
<thead>
<tr>
<th>Refinement</th>
<th>No</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CorLoc for FV+CNN</td>
<td></td>
</tr>
<tr>
<td>Horse</td>
<td>55.6</td>
<td>70.5 (+14.9)</td>
</tr>
<tr>
<td>Dog</td>
<td>37.3</td>
<td>48.4 (+11.4)</td>
</tr>
<tr>
<td>Cat</td>
<td>24.8</td>
<td>35.6 (+10.8)</td>
</tr>
</tbody>
</table>
The relation between CorLoc and detection AP

- Relation between localization during training and final test performance
  - Each of the 20 classes gives a point on the graph
Relative performance of weakly supervised learning

- Ratio of detection AP with weakly supervised training (image-labels) and AP with same detector trained from bounding box annotations
  - Each point represents one object category
Overview of the state of the art

- Methods divided into those that use external training data to learn CNN features and those that do not

<table>
<thead>
<tr>
<th></th>
<th>aero</th>
<th>bicy</th>
<th>bird</th>
<th>boa</th>
<th>bot</th>
<th>bus</th>
<th>car</th>
<th>cat</th>
<th>cha</th>
<th>cow</th>
<th>dtab</th>
<th>dog</th>
<th>hors</th>
<th>mbik</th>
<th>pers</th>
<th>plnt</th>
<th>she</th>
<th>sofa</th>
<th>trai</th>
<th>tv</th>
<th>Av.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pandey and Lazebnik’11 [32]</td>
<td>11.5</td>
<td>—</td>
<td>—</td>
<td>3.0</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>20.3</td>
<td>9.1</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>13.2</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Siva and Xiang’11 [42]</td>
<td>13.4</td>
<td>44.0</td>
<td>3.1</td>
<td>3.1</td>
<td>0.0</td>
<td>31.2</td>
<td>43.9</td>
<td>7.1</td>
<td>0.1</td>
<td>9.3</td>
<td>9.9</td>
<td>1.5</td>
<td>29.4</td>
<td>38.3</td>
<td>4.6</td>
<td>0.1</td>
<td>0.4</td>
<td>3.8</td>
<td>34.2</td>
<td>0.0</td>
<td>13.9</td>
</tr>
<tr>
<td>Russakovsky et al.’12 [35]</td>
<td>30.8</td>
<td>25.0</td>
<td>—</td>
<td>3.6</td>
<td>—</td>
<td>26.0</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>21.3</td>
<td>29.9</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>21.3</td>
<td>29.9</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Ours (FV-only)</td>
<td>36.9</td>
<td>38.3</td>
<td>11.5</td>
<td>11.1</td>
<td>1.0</td>
<td>39.8</td>
<td>45.7</td>
<td>16.5</td>
<td>1.2</td>
<td>26.4</td>
<td>4.3</td>
<td>17.7</td>
<td>31.8</td>
<td>44.0</td>
<td>13.1</td>
<td>31.4</td>
<td>9.7</td>
<td>38.5</td>
<td>36.9</td>
<td>23.3</td>
<td></td>
</tr>
</tbody>
</table>

methods using additional training data

<table>
<thead>
<tr>
<th></th>
<th>aero</th>
<th>bicy</th>
<th>bird</th>
<th>boa</th>
<th>bot</th>
<th>bus</th>
<th>car</th>
<th>cat</th>
<th>cha</th>
<th>cow</th>
<th>dtab</th>
<th>dog</th>
<th>hors</th>
<th>mbik</th>
<th>pers</th>
<th>plnt</th>
<th>she</th>
<th>sofa</th>
<th>trai</th>
<th>tv</th>
<th>Av.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Song et al.’14 [43]</td>
<td>27.6</td>
<td>41.9</td>
<td>19.7</td>
<td>9.1</td>
<td>10.4</td>
<td>35.8</td>
<td>39.1</td>
<td>33.6</td>
<td>0.6</td>
<td>20.9</td>
<td>10.0</td>
<td>27.7</td>
<td>29.4</td>
<td>39.2</td>
<td>9.1</td>
<td>19.3</td>
<td>20.5</td>
<td>17.1</td>
<td>35.6</td>
<td>7.1</td>
<td>22.7</td>
</tr>
<tr>
<td>Song et al.’14 [44]</td>
<td>36.3</td>
<td>47.6</td>
<td>23.3</td>
<td>12.3</td>
<td>11.1</td>
<td>36.0</td>
<td>46.6</td>
<td>25.4</td>
<td>0.7</td>
<td>23.5</td>
<td>12.5</td>
<td>23.5</td>
<td>27.9</td>
<td>40.9</td>
<td>14.8</td>
<td>19.2</td>
<td>24.2</td>
<td>17.1</td>
<td>37.7</td>
<td>11.6</td>
<td>24.6</td>
</tr>
<tr>
<td>Bilen et al.’14 [6]</td>
<td>42.2</td>
<td>43.9</td>
<td>23.1</td>
<td>9.2</td>
<td>12.5</td>
<td>44.9</td>
<td>45.1</td>
<td>24.9</td>
<td>8.3</td>
<td>24.0</td>
<td>13.9</td>
<td>18.6</td>
<td>31.6</td>
<td>43.6</td>
<td>7.6</td>
<td>20.9</td>
<td>26.6</td>
<td>20.6</td>
<td>35.9</td>
<td>29.6</td>
<td>26.4</td>
</tr>
<tr>
<td>Wang et al.’14 [50]</td>
<td>48.8</td>
<td>41.0</td>
<td>23.6</td>
<td>12.1</td>
<td>11.1</td>
<td>42.7</td>
<td>40.9</td>
<td>35.5</td>
<td>11.1</td>
<td>36.6</td>
<td>18.4</td>
<td>35.3</td>
<td>34.8</td>
<td>51.3</td>
<td>17.2</td>
<td>17.4</td>
<td>26.8</td>
<td>32.8</td>
<td>35.1</td>
<td>45.6</td>
<td>30.9</td>
</tr>
<tr>
<td>Wang et al.’14 [50] +context</td>
<td>48.9</td>
<td>42.3</td>
<td>26.1</td>
<td>11.3</td>
<td>11.9</td>
<td>41.3</td>
<td>40.9</td>
<td>34.7</td>
<td>10.8</td>
<td>34.7</td>
<td>18.8</td>
<td>34.4</td>
<td>35.4</td>
<td>52.7</td>
<td>19.1</td>
<td>17.4</td>
<td>35.9</td>
<td>33.3</td>
<td>34.8</td>
<td>46.5</td>
<td>31.6</td>
</tr>
<tr>
<td>Ours</td>
<td>39.3</td>
<td>43.0</td>
<td>28.8</td>
<td>20.4</td>
<td>8.0</td>
<td>45.5</td>
<td>47.9</td>
<td>22.1</td>
<td>8.4</td>
<td>33.5</td>
<td>23.6</td>
<td>29.2</td>
<td>38.5</td>
<td>47.9</td>
<td>20.3</td>
<td>20.0</td>
<td>35.8</td>
<td>30.8</td>
<td>41.0</td>
<td>20.1</td>
<td>30.2</td>
</tr>
</tbody>
</table>

Results comparable with the state of the art (with CNN features), or better when no external training data is used

A lot of improvement in performance of weakly supervised detection in recent years: AP values have doubled!
Conclusion

- Presented a state-of-the-art weakly supervised object detection method
  - Strong appearance cues for recognition: FV and CNN descriptors
  - Re-localization bias suppression: Multi-fold MIL training
  - Localization refinement: alignment with long contours

- Future directions:
  - Dealing with noise on the image labels
  - Concurrent training of categories: leverage explaining away
  - Richer interactions between recognition and segmentation
Object category localization with incomplete supervision

Jakob Verbeek
LEAR team, INRIA, Grenoble, France

Joint work with: Gokberk Cinbis and Cordelia Schmid