Weakly Supervised Object Localization Using Size Estimates

Miaojing Shi and Vittorio Ferrari

**Task**
Weakly supervised object localization
input: motorbike

Current way: training on the entire set at the same time

Our way: curriculum learning using object size estimates

- **Size Order**
  - Big (easy)
  - Small (hard)

- **Size Weighting**

![Multiple Instance Learning (MIL) Baseline](image)

+ Size Order
Ground Truth

**Method**

**RE-TRAINING**

- Size Order

![RE-LOCALIZATION](image)

**Size Weighting**

Size weighting function

\[ W(p; s_w, \sigma, \delta) = \min \left( \frac{1}{1 + e^{\delta (s_w - 3\sigma - s_p)}}, \frac{1}{1 + e^{\delta (s_p - s_w - 3\sigma)}} \right) \]

- \( s_w \): estimated object size;
- \( s_p \): window proposal size;
- \( \sigma \): stdev

**Object detection**

![Baseline](image)

Less accurate estimator only marginally affects inter-batch size order

**Results on PASCAL VOC 07 (20 classes)**

<table>
<thead>
<tr>
<th>MIL Baseline*</th>
<th>+ Size Order</th>
<th>+ Size Weighting</th>
<th>Ground Truth</th>
</tr>
</thead>
</table>

![Class – Chair](image)

* AlexNet features + Linear SVM + Objectness [Dollar ECCV14] *

- Both size order and size weight improve results
- Full system outperforms state-of-the-art

**Input:** 4096D CNN image features

- **Output:** size \( \sqrt{S_o}; \sigma \)

- **Method:** kernel ridge regression

- **Class-specific regressor**

- **Train:** PASCAL VOC 12 trainval
- **Test:** PASCAL VOC 07 trainval

**Size Estimator**

<table>
<thead>
<tr>
<th>N</th>
<th>Kendall’s ( \tau )</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td>0.614</td>
<td>0.013</td>
</tr>
<tr>
<td>100</td>
<td>0.561</td>
<td>0.016</td>
</tr>
<tr>
<td>50</td>
<td>0.542</td>
<td>0.018</td>
</tr>
<tr>
<td>40</td>
<td>0.530</td>
<td>0.019</td>
</tr>
<tr>
<td>30</td>
<td>0.527</td>
<td>0.020</td>
</tr>
</tbody>
</table>

\( N \) - number of training samples per class

**Inter-batch order**

\[
\text{recall} = \frac{|Q_{GT} \cap Q_{ES}|}{|Q_{GT}|}
\]

- \( Q_{GT} \): Ground truth sequence
- \( Q_{ES} \): Estimated sequence

**Train on trainval, measure CorLoc; Test on test, measure mAP**

<table>
<thead>
<tr>
<th>Method</th>
<th>CorLoc</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>39.1</td>
<td>20.1</td>
</tr>
<tr>
<td>Our Scheme</td>
<td>46.3</td>
<td>24.9</td>
</tr>
<tr>
<td>Baseline</td>
<td>53.8</td>
<td>28.0</td>
</tr>
<tr>
<td>Cinbis PAMI16</td>
<td>60.9</td>
<td>36.0</td>
</tr>
<tr>
<td>Wang TIP15</td>
<td>43.2</td>
<td>24.7</td>
</tr>
<tr>
<td>Blen CVP15</td>
<td>54.2</td>
<td>28.6</td>
</tr>
<tr>
<td>Dollar ECCV14</td>
<td>48.5</td>
<td>31.6</td>
</tr>
<tr>
<td>Fast R-CNN [Girshick CVPR15]</td>
<td>43.7</td>
<td>27.7</td>
</tr>
</tbody>
</table>

- Both size order and size weight improve results
- Full system outperforms state-of-the-art

**Deep v.s. Deeper**

- AlexNet [Krizhevsky NIPS12] 60.9
- VGG16 [Simonyan ICLR15] 64.7
- Baseline: 0.464

**Size estimator:** training set size

- \( N=\text{ALL} \)
- \( N=50 \)
- \( N=30 \)

**Size estimator:** generalization across classes

- class-specific
- class-generic
- across-class

**Weakly supervised object localization**

Our way: curriculum learning using object size estimates

- Current way: training on the entire set at the same time

- **Size Order**
  - Big (easy)
  - Small (hard)

- **Size Weighting**

Input: 4096D CNN image features

Output: size

Method: kernel ridge regression

Class-specific regressor

Train: PASCAL VOC 12 trainval

Test: PASCAL VOC 07 trainval

\( \sigma \): stdev

**Deep v.s. Deeper**

- AlexNet [Krizhevsky NIPS12] 60.9
- VGG16 [Simonyan ICLR15] 64.7

AlexNet features + Linear SVM + Objectness [Dollar ECCV14] *