Learning Semantic Segmentation with Weakly-Annotated Videos

Pavel Tokmakov  
Karteek Alahari  
Cordelia Schmid

Inria, France

Goals
- Train a model for semantic segmentation
- Use as little supervision as possible

Overview
- Our approach uses motion as a cue in label inference
- Robust to errors in motion segmentation methods [3]

State-of-the-art
- Based on FCNN framework [Chen et al., ICLR 2015]
- EM-like approaches: E-step - pixel label inference from image tags, M-step - backpropagation
- Label inference requires manual constraints on the size of segments [1, 2] and lacks precision

Our approach
Integrates motion and semantic cues

Experimental evaluation
Comparison to [1]

<table>
<thead>
<tr>
<th>Method</th>
<th>Trained on</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>YouTube</td>
<td>23.3</td>
<td></td>
</tr>
<tr>
<td>VOC</td>
<td>35.8</td>
<td></td>
</tr>
<tr>
<td>VOC aug.</td>
<td>40.2</td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>41.2</td>
<td></td>
</tr>
</tbody>
</table>

Comparison to the state-of-the-art

<table>
<thead>
<tr>
<th>Method</th>
<th>Training data</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strong/Full supervision [1]</td>
<td>YouTube, VOC, VOC aug.</td>
<td>40.2</td>
</tr>
<tr>
<td>Weak supervision with additional info [2]</td>
<td>YouTube, VOC, VOC aug.</td>
<td>39.0</td>
</tr>
<tr>
<td>Ours</td>
<td>YouTube, VOC, VOC aug.</td>
<td>43.7</td>
</tr>
</tbody>
</table>

Qualitative results

Co-localization
- Replace location unary in [3] with our predictions
- Evaluate co-localization performance on YouTubeObjects

Training
Datasets
- YouTubeObjects: 10 categories, 155 videos, 2511 shots
- ImageNet videos: 10 categories, 795 videos, 2120 shots
- Pascal VOC 2012: 20 categories (10 corresponding to YouTubeObjects used on validation set), 1469 images

Training procedure
- Prune shots with erroneous motion segmentations based on segment size
- Train our model on the remaining shots
- Finetune on small subset, selecting one shot per video with previously trained model
- When training with images and videos jointly use [1] for label inference on images

References