Joint learning of object and action detectors

Vicky Kalogeiton\textsuperscript{1,2}, Philippe Weinzaepfel\textsuperscript{3}, Vittorio Ferrari\textsuperscript{2} and Cordelia Schmid\textsuperscript{1}

Introduction

Joint object-action detection in videos

Related work

Objects or actions

[Image 31x71 to 302x116]

[Image 46x1074 to 895x1407]

Contributions

- End-to-end multitask architecture
- Enables zero-shot learning, relationship detection in images and object-action semantic segmentation
- Fewer parameters than alternatives:
  Multitask
  Hierarchical

Zero-shot learning

Remove action labels for one object category at training time but predict them at test time

\[ \text{Ours: SharpMask} \]

Relationship detection in images

- VDO dataset: 288 object relationships in triplets: object-interaction-object
- Transform each triplet into two pairs: object and an interaction label

Qualitative results

Mean IoU segmentation accuracy

Quantitative results

Global pixel accuracy

Joint object-action detection results

End-to-end network

Joint learning

<table>
<thead>
<tr>
<th>A2D</th>
<th>YTO</th>
<th>ILSVRC VID</th>
</tr>
</thead>
<tbody>
<tr>
<td>objects</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>actions</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A2D</th>
<th>YTO</th>
<th>ILSVRC VID</th>
</tr>
</thead>
<tbody>
<tr>
<td>objects mAP</td>
<td>69.6%</td>
<td>66.9%</td>
</tr>
<tr>
<td>actions mAP</td>
<td>59.0%</td>
<td>53.2%</td>
</tr>
</tbody>
</table>

Better fusion ✓ Joint better than separate learning ✓ Less parameters

Qualitative results

Frame

Ground truth

Ours