Online Real-time Multiple Spatiotemporal Action Localisation and Prediction

Gurkirt Singh¹  Suman Saha¹  Michael Sapienza²  Philip Torr²  Fabio Cuzzolin¹
¹Oxford Brookes University  ²University of Oxford

Online Action Localisation and Prediction

Task is to determine what action is occurring in a video, as early as possible and localise it.

- **Action Localisation**: defined as a set of linked bounding boxes covering each individual action instance, called action tubes.
- **Online**: method designed to construct tubes incrementally.
- **Prediction**: to predict the label of the video at any given point of time, for e.g. when only 10% of the video has been observed.

Why?

- Real-time online action localisation is essential for many applications, for e.g. surveillance, human-robot interaction.
- Early action label prediction is crucial in intervention applications, for e.g. surgical robotics or autonomous driving.

Contributions

- Unlike previous methods [1,2,3], we construct multiple action tubes simultaneously and incrementally, getting rid of recursive calls to dynamic programming to generate multiple tubes.
- First method to perform online spatiotemporal action localisation in real-time, while still outperforming previous methods.
- Unlike [4], we perform early action localisation and prediction in untrimmed videos of the UCF101 dataset.

Early label prediction

- Early label prediction is a side product of our online tube generation algorithm.
- At any given point of time the video is assigned the label of the tube with highest score from the current set of tubes.

Online and Real-Time Pipeline

- We provide the option of using two different type of optical flows as shown in above figure in block (b) and (c).
- Two SSD [7] networks (e) are used for appearance & optical flow detection.
- We provide two options for fusing two sets of output detections (f) from appearance and flow networks (g).
- Tube generation (h) is performed for each class c in parallel and independently, frame by frame starting from first frame.

Step-1: Initialise the tubes using top n detections for class c.

Step-2: At any time t sort the tubes based on their score.

Step-3: For each tube find potential matches based on the IoU and the box scores for class c as illustrated below:

- Score: 0.6 tube id: 1 tube index: 1
- Score: 0.7 tube id: 2 tube index: 1
- Score: 0.6 tube id: 1 tube index: 2
- Score: 0.9 tube id: 2 tube index: 2

Step-4: Update temporal labelling of each tube using the new added box, by using a temporal label changing cost as shown by [8].

Step-5: Terminate tubes based on the labeling or if no match is found for the tube in the past k frames.

Step-6: Initialise new tubes using unassigned detections.

- We used n = 10, λ = 0.1 and k = 5 in all of our experiments.

Results

Early Label Prediction

Online localisation-JHMDB21

<table>
<thead>
<tr>
<th>Method</th>
<th>threshold</th>
<th>0.2</th>
<th>0.5</th>
<th>0.75</th>
<th>0.5/0.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peng et al. [5]</td>
<td>73.5</td>
<td>32.1</td>
<td>02.7</td>
<td>07.3</td>
<td></td>
</tr>
<tr>
<td>Saha et al. [6]</td>
<td>66.6</td>
<td>36.4</td>
<td>07.9</td>
<td>14.4</td>
<td></td>
</tr>
</tbody>
</table>

Spatiotemporal localisation results on UCF101*

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Peng et al. [5]</td>
<td>69.8</td>
<td>40.9</td>
<td>15.5</td>
<td>18.7</td>
<td>70.2</td>
<td>43.0</td>
<td>14.5</td>
</tr>
<tr>
<td>Saha et al. [6]</td>
<td>73.0</td>
<td>44.0</td>
<td>14.1</td>
<td>19.2</td>
<td>75.3</td>
<td>46.3</td>
<td>15.0</td>
</tr>
</tbody>
</table>

*R: All the result are shown on revised spatiotemporal annotations, corrected manually by authors

Real-time analysis

<table>
<thead>
<tr>
<th>Modules</th>
<th>Setup</th>
<th>Flow computation time (ms)</th>
<th>Detection network time (ms)</th>
<th>Tube generation time (ms)</th>
<th>Overall speed (fps)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Module 1</td>
<td>A</td>
<td>07.0</td>
<td>21.8</td>
<td>02.5</td>
<td>40.0</td>
</tr>
<tr>
<td>Module 2</td>
<td>A+RTF</td>
<td>110</td>
<td>145</td>
<td>10.0</td>
<td>70.0</td>
</tr>
<tr>
<td>Module 3</td>
<td>A+AF</td>
<td>03.0</td>
<td>145</td>
<td>10.0</td>
<td>80.0</td>
</tr>
</tbody>
</table>

References


Code: https://github.com/gurkirt/online-action-detection
Email: Gurkirt.Singh@brookes.ac.uk
Revised annotations: https://github.com/gurkirt/corrected-UCF101-Annots