Joint Tracking and Segmentation of Multiple Targets

Anton Milan¹, Laura Leal-Taixé², Konrad Schindler², Ian Reid¹
¹University of Adelaide, Australia. ²Photogrammetry and Remote Sensing Group, ETH Zürich.

Figure 1: Qualitative tracking and segmentation results.

Tracking-by-detection traditionally relies on a set of sparse detections that serve as input to a high-level tracker whose goal is to correctly associate these “dots” over time. An obvious shortcoming of this approach is that most information available in image sequences is simply ignored by thresholding weak detection responses and applying non-maximum suppression. We argue that it is beneficial to consider all image evidence to handle tracking in crowded scenarios. In contrast to many previous approaches, we aim to assign a unique target ID not only to each individual detection, but to every (super-)pixel in the entire video (cf. Fig. 1). This low-level information enables us to recover trajectories of largely occluded targets since the partial image evidence of the superpixels often persists even in the absence of detections. Exploiting low-level information in the context of multi-target tracking has been recently proposed [2, 3]. However, one major limitation of previous approaches is their inherent inability to track targets through full occlusions. In addition, a target’s state (i.e. its location) is only defined implicitly by the segmentation [2], which makes it rather difficult to estimate the full extent in case of (partial) occlusion. Our experiments confirm that these methods show relatively poor performance in crowded scenes when evaluated with standard multi-target tracking measures. This work overcomes both limitations by explicitly modelling the continuous state of all targets throughout the entire sequence.

In common with some other approaches [5–7], we formulate the problem as one of finding a set of continuous trajectory hypotheses that best explains the data, but our approach differs in that we take account of the low-level information in scoring the trajectory hypotheses. We do this by modelling the problem as a multi-label conditional random field (CRF). Furthermore, contrary to prior closely related work [1, 6], our trajectory model is not a simple space-time curve but rather a volumetric tube with a rectangular cross-section, allowing for a more accurate representation. Our method shows encouraging results on many standard benchmark sequences and significantly outperforms state-of-the-art tracking-by-detection approaches in crowded scenarios. In contrast to many previous approaches, we aim our joint tracking and segmentation framework provides reasonable instance-based segmentation masks in crowded scenarios.

Our high-level approach to this problem follows a model selection strategy, similar to [1]: we generate an overcomplete set of trajectory hypotheses to capture agreement with image evidence. We do this by involving the pixel information in the optimization we enable the label IDs to persist even when there is no explicit detector evidence. Tab. 1 shows results of our approach compared to two top public submissions on the recent MOTChallenge benchmark. Further model details and more experiments can be found in the paper.

Our conclusion is that exploiting all image information helps to improve multiple target tracking in regions of long-term partial occlusions. Moreover, our joint tracking and segmentation framework provides reasonable instance-based segmentation masks in crowded scenarios.

Table 1: Results on the MOTChallenge 2015 Benchmark.

<table>
<thead>
<tr>
<th>Method</th>
<th>TA</th>
<th>TP</th>
<th>Rcll</th>
<th>Prcl</th>
<th>MT</th>
<th>ML</th>
<th>ID</th>
<th>FM</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMOT [8]</td>
<td>18.6</td>
<td>69.6</td>
<td>40.0</td>
<td>66.4</td>
<td>5.3</td>
<td>53.3</td>
<td>684</td>
<td>1282</td>
</tr>
<tr>
<td>CEM [6]</td>
<td>19.3</td>
<td>70.7</td>
<td>43.7</td>
<td>65.4</td>
<td>8.5</td>
<td>46.5</td>
<td>813</td>
<td>1023</td>
</tr>
<tr>
<td>MotCon [4]</td>
<td>23.1</td>
<td>70.9</td>
<td>41.7</td>
<td>71.1</td>
<td>4.7</td>
<td>52.0</td>
<td>1018</td>
<td>1061</td>
</tr>
<tr>
<td>SegTrack</td>
<td>22.5</td>
<td>71.7</td>
<td>36.5</td>
<td>74.0</td>
<td>5.8</td>
<td>63.9</td>
<td>697</td>
<td>737</td>
</tr>
</tbody>
</table>