1. Classification by Conv-LSTM, which can preserve spatial-temporal information
2. Various post-processing to suppress false alarm
3. Proposal alignment to learn efficiently with few training data

System overview

Input → Object detection → Proposal generation → Feature extraction → Classification → Post-processing → Event category and location

- **Object detection**: detect objects by Mask R-CNN fine-tuned with our extra annotation
- **Proposal generation**: track each object by tracking method and merge them if person and vehicle are nearby
- **Feature extraction**: divide proposal into snippets and extract features using I3D network

Feature extraction and classification

- X and y axis of I3D feature have spatial information, and z axis has temporal information
- Problem: simple flattening algorithm loses spatial-temporal information
- Solution: reshape I3D features and use Conv-LSTM

Flow of action likelihood prediction for each clip

Proposal → for each clip → 64 frame snippet → I3D → 5-th layer → 7x7x8 dim. → 1024 ch. → Conv-LSTM → Likelihood

Proposal alignment*

- Observation: Each action has diversity of appearance due to various movement/object direction
- Assumption: this diversity makes learning and predicting action recognition difficult
- Solution: rotate proposals to align movement/object direction

Post-processing

- Problem: simply merging results of each clip produces extra estimation results
- Solution: remove extra results by likelihood consistency check, threshold, merge, label consistency check and NMS

Proposal alignment*  * it's not included in submitted system

Evaluation

Whole system results on test data
- 5th accuracy at time of submission

Each module results on validation data

Object detection results

<table>
<thead>
<tr>
<th>Method</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original Mask R-CNN</td>
<td>19.6%</td>
</tr>
<tr>
<td>Fine-tuned Mask R-CNN</td>
<td>44.1%</td>
</tr>
</tbody>
</table>

Proposal generation results

<table>
<thead>
<tr>
<th>Method</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>85.6%</td>
</tr>
</tbody>
</table>

Classification results
- We used our proposals

Proposal alignment results
- We used GT proposals

<table>
<thead>
<tr>
<th>Method</th>
<th>Number of proposals</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>4,151</td>
<td>85.6%</td>
</tr>
<tr>
<td>With alignment</td>
<td>46.1%</td>
<td>49.7%</td>
</tr>
<tr>
<td>Without alignment</td>
<td>13.1%</td>
<td>49.7%</td>
</tr>
</tbody>
</table>