Human Intention Prediction Using Two-Stream Spatio-Temporal Features

Shengchao Li, Lin Zhang, and Xiumin Diao
School of Engineering Technology, Purdue University
Phone: +1 (765) 494-2212; Email: diaox@purdue.edu

ACM/IEEE International Conference on Human-Robot Interactions
March 11-14, 2018, Daegu, Korea
The experiment is about a human participant pitching a ball toward a robot (represented by a target area in the figure). The target area is divided into 9 grids which represent the 9 targets that the ball can hit. We expect “the robot” to be able to predict the intention of the participant, namely, which of the nine targets the participant is pitching the ball to.
Training Data

RGB images sequence

Optical flow sequence
Spatial network:
Optimize early fusion methods of RGB images and optical flow for spatial network.

Three channels of RGB images

Three channels of RGB images + Optical flow

One channel of Optical RGB images + flow

Spatial network

Loss of spatial network
Approach

Temporal network:
Capture motion information in temporal network.

Optical flow sequence of a video

Subsampling

Temporal network

Loss of temporal network
Approach

Two-stream network with average fusion: Final average fusion to improve prediction performance.

RGB images + Optical flow

Spatial network

Prediction

Average fusion

Temporal network

Prediction
## Results

<table>
<thead>
<tr>
<th>Spatial Network Method</th>
<th>Data Augmentation</th>
<th>Prediction Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Three channels of RGB images</td>
<td>No</td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>57.4%</td>
</tr>
<tr>
<td>Three channels of RGB images + two channels of optical flow</td>
<td>No</td>
<td>53%</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>57.4%</td>
</tr>
<tr>
<td>One channel of RGB images + two channels of optical flow</td>
<td>No</td>
<td>51.8%</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>63.89%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Temporal Network Method</th>
<th>Data Augmentation</th>
<th>Prediction Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temporal network without fusion</td>
<td>No</td>
<td>68.5%</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>69.4%</td>
</tr>
<tr>
<td>Final fusion of spatial network and temporal network</td>
<td>Yes</td>
<td>71.3%</td>
</tr>
</tbody>
</table>
Thank you!