Human activity recognition

Group activity recognition

• Real video from a CCTV in Malaysia
  ▪ A group of thieves steal a laptop by distracting the owner

Ryoo & Aggarwal, IJCV 2011
Aerial video understanding

Recognition of human activities from UAVs
Robot-centric activity recognition

Providing activity-level situation awareness to robots

• In real-world robot scenarios, a robot is expected to face various types of human activities.

Human actions

Human interactions

First-person activities

[ Laptev, IJCV 2005 ]

[ Ryoo and Aggarwal, ICCV 2009 ]

[ Ryoo and Matthies, CVPR 2013 ]

Recognition of each type has been studied separately

• Our objective: investigation of a unified approach
First-person activity recognition
End-to-end activity learning

Deep learning of activity temporal structure (e.g., sub-events)

- We propose a new approach of learning ‘temporal’ attention filters on top of convolutional neural networks

```
Raw video input

Per-frame CNNs

Temporal attention filters

FC Layer

FC Layer

\{v_1, v_2, \ldots, v_d\}

\{v_1, v_2, \ldots, v_d\}

\{v_1, v_2, \ldots, v_d\}
```

6
Multi-type activity recognition

[ICRA 2016 best vision paper]

Video frames and estimated skeletons
Experimental results

Evaluated our approach with 4 different datasets

- Human-human, 1\textsuperscript{st}-person, human action, and multi-type.

<table>
<thead>
<tr>
<th>Method</th>
<th>SBU</th>
<th>FP</th>
<th>MSR</th>
<th>MT-Acc</th>
<th>MT-F1</th>
<th>MT-F2</th>
<th>Real-time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yun et al. [21]</td>
<td>80.03%</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Y</td>
</tr>
<tr>
<td>Ji et al. [30]</td>
<td>86.9%</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Y</td>
</tr>
<tr>
<td>Oreifej et al. [31]</td>
<td>77.0%</td>
<td>45.55%</td>
<td>88.36%</td>
<td>68.07%</td>
<td>0.7315</td>
<td>0.7870</td>
<td>N</td>
</tr>
<tr>
<td>Xia et al. [8]</td>
<td>42.69%</td>
<td>53.25%</td>
<td>37.76%</td>
<td>28.38%</td>
<td>0.3891</td>
<td>0.46</td>
<td>N</td>
</tr>
<tr>
<td>Laptev et al. [1]</td>
<td>66.28%</td>
<td>50.84%</td>
<td>-</td>
<td>38.59%</td>
<td>0.4519</td>
<td>0.5641</td>
<td>N</td>
</tr>
<tr>
<td>Xia et al. [32]</td>
<td>41.88%</td>
<td>70.0%</td>
<td>78.97%</td>
<td>47.06%</td>
<td>0.4434</td>
<td>0.5075</td>
<td>Y</td>
</tr>
<tr>
<td>Xia et al. [7]</td>
<td>-</td>
<td>83.7%</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>N</td>
</tr>
<tr>
<td>Li et al. [22]</td>
<td>-</td>
<td>-</td>
<td>74.7%</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>?</td>
</tr>
<tr>
<td>Wang et al. [33]</td>
<td>-</td>
<td>-</td>
<td>86.5%</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>?</td>
</tr>
<tr>
<td>Wang et al. [10]</td>
<td>-</td>
<td>-</td>
<td>88.2%</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>?</td>
</tr>
<tr>
<td>Wang et al. [24]</td>
<td>-</td>
<td>-</td>
<td>90.22%</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>?</td>
</tr>
<tr>
<td>Evangelidis et al. [34]</td>
<td>-</td>
<td>-</td>
<td>89.86%</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>?</td>
</tr>
<tr>
<td>Chaaraoui et al. [35]</td>
<td>-</td>
<td>-</td>
<td>91.8%</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Y</td>
</tr>
<tr>
<td>Vemulapalli et al. [26]</td>
<td>-</td>
<td>-</td>
<td>92.46%</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>?</td>
</tr>
<tr>
<td>Luo et al. [25]</td>
<td>-</td>
<td>-</td>
<td>96.70%</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>?</td>
</tr>
<tr>
<td><strong>RHI (ours)</strong></td>
<td><strong>93.08%</strong></td>
<td><strong>85.94%</strong></td>
<td><strong>95.38%</strong></td>
<td><strong>76.67%</strong></td>
<td><strong>0.7954</strong></td>
<td><strong>0.8633</strong></td>
<td><strong>Y</strong></td>
</tr>
</tbody>
</table>
Ongoing: robot learning to replicate activities

Learning ‘actionable’ representations of human activities

- Representations not for the *recognition* but for the robot *execution* of the learned activities.

Video examples

Learned representation

Robot execution
THANK YOU

CONTACT INFORMATION

mryoo@indiana.edu
http://michaelryoo.com