Action Assessment by Joint Relation Graphs

Jia-Hui Pan 1, Jibin Gao 1, Wei-Shi Zheng 1,2,3

1 Sun Yat-sen University, China
2 Peng Cheng Laborator
3 Key Laboratory of Machine Intelligence and Advanced Computing

ICCV 2019
**Introduction**

**Action Assessment:** Video -> Score

- **Whole Scene**  
  Prior works

- **Joint Relations**  
  This paper

- **Separate Joints**  
  Prior works

**Why joint relations?** An example in diving:

- Bending knee + bending ankle and hip = Good (e.g. the rolling stage)
- Bending knee + straight ankle and hip = Bad (e.g. the water-entering stage)
Introduction

Good Performance

Body Part Movement  Body Coordination
(The commonality module)  (The difference module)

Spatial Relation Graph  Temporal Relation Graph

Assessing the Performance of an Action

The Commonality: Body Part Kinetics
The Difference: Joint Coordination

Good Performance = Excellent Movement for each body part + Good Coordination among joints
Overview
Spatial Relation Graph

**Nodes:** Joints at the same time step

**Edge:** Learnable relations between joints

\[
A_s \in \mathbb{R}^{J \times J}
\]

Non-negative and learnable

Edges not in the skeleton are set as zero

Temporal Relation Graph

**Nodes:** Joints at the adjacent time step

**Edge:** Learnable relations between joints

\[
A_p \in \mathbb{R}^{J \times J}
\]

Non-negative and learnable
The Commonality Module

Learning the features within locally connected joints

Updated features \( H_1^t \) = \( A_s \cdot H_0^t \),

\( H_0^t, H_1^t \in \mathbb{R}^{J \times M} \)

Feature aggregation by average pooling:

\( \bar{h}_c^t = \frac{1}{N}(H_c^{tT} \cdot 1) \),
The Difference Module

Learning coordination in joint neighbourhoods

Spatial difference: $J \times M$

Extracted video features

$$D_s^t(i, m) = \sum_j (A_s(i, j) \cdot (F^t(i, m) - F^t(j, m))) \cdot w_j,$$

$$D_p^t(i, m) = \sum_j (A_p(i, j) \cdot (F^t(i, m) - F^{t-1}(j, m))) \cdot w_j,$$

$1 \leq i, j \leq J, 1 \leq m \leq M,$

Temporal difference: $J \times M$

Learnable weight

Feature aggregation by average pooling:

$$\bar{d_s}^t = \frac{1}{N} (D_s^T \cdot 1),$$
Regression Module

Input:
The whole scene feature \( q^t \)
The commonality features \( \tilde{h}_0^t, \tilde{h}_1^t \)
The difference features \( \tilde{d}_s^t \) and \( \tilde{d}_p^t \)

Weighted feature pooling:

\[
v^t = \sum_i \alpha_i \cdot \hat{u}_i^t + \beta_i
\]

Orthogonal regularization:

\[
R_O = \sum_{i,j} \gamma \cdot (\hat{u}_i^T \cdot \hat{u}_j^t)
\]

Final regression with two FCs:

\[
s = \sum_t S(v^t)
\]
Experiments

I3D features and Mask-RCNN human poses

The Olympic Actions

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ST-GCN [33]</td>
<td>0.5300</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C3D-LSTM [21]</td>
<td>0.3286</td>
<td>0.5770</td>
<td></td>
<td>0.1681</td>
<td>0.1234</td>
<td>0.6600</td>
</tr>
<tr>
<td>C3D-SVR [21]</td>
<td>0.6047</td>
<td>0.5636</td>
<td>0.4593</td>
<td>0.5029</td>
<td>0.7912</td>
<td>0.6927</td>
</tr>
<tr>
<td>Whole Scene</td>
<td>0.6339</td>
<td>0.6872</td>
<td>0.5179</td>
<td>0.5053</td>
<td>0.8783</td>
<td>0.8832</td>
</tr>
<tr>
<td>Whole+Patch</td>
<td>0.7043</td>
<td>0.6758</td>
<td>0.5783</td>
<td>0.4547</td>
<td>0.8547</td>
<td>0.8766</td>
</tr>
<tr>
<td>Ours</td>
<td>0.7630</td>
<td>0.7358</td>
<td>0.6006</td>
<td>0.5405</td>
<td>0.9013</td>
<td>0.9254</td>
</tr>
</tbody>
</table>

Skeleton on JIGSAWS

JIGSAWS: Kinematic features for joints

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ST-GCN [33]</td>
<td>0.31</td>
<td>0.39</td>
<td>0.58</td>
<td>0.43</td>
</tr>
<tr>
<td>TSN [6]</td>
<td>0.34</td>
<td>0.23</td>
<td>0.72</td>
<td>0.46</td>
</tr>
<tr>
<td>Whole-scene</td>
<td>0.09</td>
<td>0.10</td>
<td>0.15</td>
<td>0.11</td>
</tr>
<tr>
<td>Joint Motion</td>
<td>0.17</td>
<td>0.25</td>
<td>0.55</td>
<td>0.34</td>
</tr>
<tr>
<td>Whole + Joint</td>
<td>0.17</td>
<td>0.37</td>
<td>0.73</td>
<td>0.46</td>
</tr>
<tr>
<td>Ours</td>
<td>0.36</td>
<td>0.54</td>
<td>0.75</td>
<td>0.57</td>
</tr>
</tbody>
</table>

Skeleton on JIGSAWS

Master Left

Master Right

Slave Left

Slave Right
<table>
<thead>
<tr>
<th></th>
<th>Diving</th>
<th>Gymvault</th>
<th>Skiing</th>
<th>Snowboard</th>
<th>Sync. 3m</th>
<th>Sync. 10m</th>
<th>Avg. Corr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours(Full)</td>
<td><strong>0.7630</strong></td>
<td><strong>0.7358</strong></td>
<td>0.6006</td>
<td><strong>0.5405</strong></td>
<td><strong>0.9013</strong></td>
<td><strong>0.9254</strong></td>
<td><strong>0.7849</strong></td>
</tr>
<tr>
<td>w/o Commonality</td>
<td>0.7020</td>
<td>0.7166</td>
<td>0.5222</td>
<td>0.5117</td>
<td>0.8632</td>
<td>0.9073</td>
<td>0.7423</td>
</tr>
<tr>
<td>w/o Difference</td>
<td>0.7469</td>
<td>0.7007</td>
<td><strong>0.6191</strong></td>
<td>0.4968</td>
<td>0.8651</td>
<td>0.8764</td>
<td>0.7455</td>
</tr>
<tr>
<td>w/o Spatial Relation</td>
<td>0.7193</td>
<td>0.6512</td>
<td>0.5059</td>
<td>0.4752</td>
<td>0.8725</td>
<td>0.8813</td>
<td>0.7229</td>
</tr>
<tr>
<td>w/o Temporal Relation</td>
<td>0.7273</td>
<td>0.6490</td>
<td>0.5186</td>
<td>0.5203</td>
<td>0.8824</td>
<td>0.9049</td>
<td>0.7423</td>
</tr>
<tr>
<td>w/o Feature Pooling</td>
<td>0.7288</td>
<td>0.7349</td>
<td>0.5504</td>
<td>0.4528</td>
<td>0.8640</td>
<td>0.9032</td>
<td>0.7451</td>
</tr>
<tr>
<td>w/o Feature Encoders</td>
<td>0.6504</td>
<td>0.6755</td>
<td>0.3088</td>
<td>0.3293</td>
<td>0.8421</td>
<td>0.8268</td>
<td>0.6512</td>
</tr>
<tr>
<td>Whole-scene (Baseline)</td>
<td>0.6339</td>
<td>0.6872</td>
<td>0.5179</td>
<td>0.5053</td>
<td>0.8783</td>
<td>0.8832</td>
<td>0.7226</td>
</tr>
</tbody>
</table>
For gymvault:

*In the spatial graph, hips, shoulders, and knees are closely related.*

*In the temporal graph, shoulders are more attended.*
Conclusion:
• Assess the action performance through graph-based joint relation modelling
• Joint commonality module and the joint difference module

Comments:
- Similar to the methods in skeleton-based action recognition
- Depend on the existence of joints and the human pose estimation method