DeepPEAR: Deep Pose Estimation and Action Recognition

Authors: You-Ying Jhuang, Wen-Jiin Tsai
National Chiao Tung University, TAIWAN

Idea
To use skeleton and appearance features to do action recognition, but use RGB video as the only input.

Main Contributions
• Propose a method to recognize actions from the predicted 3D pose and the appearance features generated by the pose estimation network
• Require less equipment, compared to the skeleton based action recognition
• Achieve state-of-the-art results on NTU RGB+D dataset
Overall Architecture

**Pose Module**

Main components:
1. Pose module
2. Skeleton module
3. Appearance module
4. Aggregation module
Pose module

- It uses stacked hourglasses networks to estimate 2D heatmaps and then uses conv. features and heatmaps to estimate depths for each body joint.
- It takes RGB video as input and outputs 3D skeleton coordinates, conv. features and 2D joint heatmap.
Skeleton module

- Skeletal data has the benefits of insensitive to illumination changes and cluttered background, and is more correlated to human actions.

*Change channel dimension from coordinates to joints*
**Appearance module**

- It takes convolutional features and body joint heatmaps as the input to predict actions.

**Kronecker product**

To extract appearance features around joints.

SE block (Squeeze-and-Excitation):
learn the weights of each channel.
Aggregation module

- It takes the prediction result from skeleton module and appearance module to produce the final prediction.
- Element-wise summation
- Element-wise multiplication
- Concatenation
  - uses convolution layers to extract fused features and predict the final classification result by fully connected layers.

<table>
<thead>
<tr>
<th>Aggregation methods</th>
<th>Acc. CS</th>
<th>Acc. CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Element-wise summation</td>
<td>91.76</td>
<td>95.25</td>
</tr>
<tr>
<td>Element-wise multiplication</td>
<td>91.53</td>
<td>95.41</td>
</tr>
<tr>
<td>Concatenation</td>
<td>88.86</td>
<td>95.06</td>
</tr>
</tbody>
</table>

Acc. : accuracy
CS: cross subject
CV: cross view
## Experiment Result

### Comparison with state-of-the-art

<table>
<thead>
<tr>
<th>Methods</th>
<th>ACC. CS</th>
<th>ACC. CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-CNN+MLTN [21] (S)</td>
<td>79.57</td>
<td>84.83</td>
</tr>
<tr>
<td>VA-LSTM [15] (S)</td>
<td>79.4</td>
<td>87.6</td>
</tr>
<tr>
<td>ST-GCN [6] (S)</td>
<td>81.5</td>
<td>88.3</td>
</tr>
<tr>
<td>SR-TSL [7] (S)</td>
<td>84.8</td>
<td>92.4</td>
</tr>
<tr>
<td>HCN [12] (S)</td>
<td>86.5</td>
<td>91.1</td>
</tr>
<tr>
<td>2D-3D-Softargmax [16] (RGB)</td>
<td>85.5</td>
<td>-</td>
</tr>
<tr>
<td>Glimpse Clouds [19] (RGB)</td>
<td>86.6</td>
<td>93.2</td>
</tr>
<tr>
<td>PoseMap [20] (RGB)</td>
<td>91.71</td>
<td>95.26</td>
</tr>
<tr>
<td><strong>Ours (RGB)</strong></td>
<td><strong>91.76</strong></td>
<td><strong>95.41</strong></td>
</tr>
</tbody>
</table>

Acc. : accuracy  
CS: cross subject  
CV: cross view
Thanks for your attention