**Human Action Recognition: Pose-based Attention draws focus to Hands**

Fabien Baradel*, Christian Wolf*, Julien Mille**

* Univ Lyon, INSA-Lyon, CNRS, LIRIS, F-69621, Villeurbanne, France
** Laboratoire d’Informatique de l’Université de Tours (EA 6300), INSA Centre Val de Loire, 41034 Blois, France

Email: fabien.baradel@liris.cnrs.fr

### PROBLEM DEFINITION & MOTIVATIONS

**Overview**
- Video Understanding
- Human Action Recognition
- Video captured by Microsoft Kinect3D
  - (3D human pose - RGB - Depth)

**Main challenges**
- High dimensional data
- Spatio-Temporal information
- Noise in the human pose
- Conditioned on augmented motion

**Problem statement:** How can an attention mechanism select the most discriminative parts of the video?

### MAIN IDEA

- **Two modalities**
  - 2D skeleton coordinates
  - RGB frames
- **Two stream model**
  - RGB
  - Spatial attention mechanism over RGB hands crops
  - Spatial attention adjusted at each timestep
  - Conditioned on augmented pose
  - Temporal Attention on hidden states

### PROPOSED APPROACH

- **STA-HANDS**
  - Spatial Attention around Hands crops
  - Inception features from RGB crops around hands
  - Attention weights computed given augmented pose

- **SA-Hands**
  - Spatial Attention around hands crops
  - Inception features from RGB crops around hands
  - Attention weights computed given augmented pose
  - Fully differentiable

- **Temporal Attention on LSTM features**
  - Can be seen as a dynamic pooling
  - Weighted average of hidden states
  - Given augmented motion
  - Fully differentiable

### EXPERIMENTAL RESULTS

**Comparison**
- State of the art on NTU RGB+D (NTU) (~57,000 videos - 60 classes)
  - First to combine 3D skeleton data and RGB frames on NTU

**Ablation Study**
- Attention Conditioning: pose features > hidden state
  - Attention mechanism has a high impact on RGB only stream
  - Augmented pose
  - Temporal Attention: + 3.2 points
  - Spatio-Temporal Attention: + 5.4 points

**Table 2:** Effects of the combination on the spatial and temporal attention (RGB stream only, accuracy in %).

<table>
<thead>
<tr>
<th>Methods</th>
<th>Spatial Attention</th>
<th>Temporal Attention</th>
<th>CS</th>
<th>CV</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>SA Hands</td>
<td></td>
<td></td>
<td>70.1</td>
<td>78.0</td>
<td>74.0</td>
</tr>
<tr>
<td>STA Hands</td>
<td></td>
<td></td>
<td>71.0</td>
<td>78.5</td>
<td>74.2</td>
</tr>
<tr>
<td>STA Hands</td>
<td></td>
<td></td>
<td>73.3</td>
<td>80.2</td>
<td>76.7</td>
</tr>
<tr>
<td>STA Hands</td>
<td></td>
<td></td>
<td>73.6</td>
<td>80.7</td>
<td>76.9</td>
</tr>
<tr>
<td>STA Hands</td>
<td></td>
<td></td>
<td>73.8</td>
<td>80.7</td>
<td>77.0</td>
</tr>
<tr>
<td>STA Hands</td>
<td></td>
<td></td>
<td>73.8</td>
<td>80.7</td>
<td>77.0</td>
</tr>
</tbody>
</table>

**Table 1:** Results on the NTU RGB-D dataset with Cross-Subject (CS) and Cross-View (CV) settings (accuracies in %, *a* means that pose is only used for the attention mechanism).