What and How? Jointly Forecasting Human Action and Pose

Yanjun Zhu, Yanxia Zhang, Qiong Liu, Andreas Girgensohn, David Doermann
Why Forecasting?

- Human daily life
- Quick response
- Imagination

Why Action and Pose Together?

- Healthcare robot
- **WHAT** is the future action and **HOW** will people perform it?
- Pose keypoints

Why Joint Learning?

- Multi-task learning
- Category classification and pose keypoints regression
- The two tasks would help each other.
Model Architecture

Encoder
Encode past frames
Recognize current action

Decoder
Predict future actions and poses

GRU
GRU
GRU

Encoder state

\(\hat{a}_t\)

\(\hat{a}_{t+1}\)
\(\hat{a}_{t+2}\)

\(\hat{p}_{t+1}\)
\(\hat{p}_{t+2}\)

\(\hat{p}_t\)
## Results: Action

<table>
<thead>
<tr>
<th>Forecasting Methods</th>
<th>( Acc_0 )</th>
<th>( Acc_1 )</th>
<th>( Acc_4 )</th>
<th>avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two-stream + FC</td>
<td>-</td>
<td>22.37</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Two-stream + LSTM</td>
<td>-</td>
<td>55.60</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>iDT + LSTM</td>
<td>-</td>
<td>65.20</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Multi-label LSTM (ours)</td>
<td>74.95</td>
<td>74.74</td>
<td>74.56</td>
<td>74.77</td>
</tr>
<tr>
<td>Multi-task Seq2seq (ours)</td>
<td><strong>76.36</strong></td>
<td><strong>76.12</strong></td>
<td>74.34</td>
<td><strong>75.21</strong></td>
</tr>
</tbody>
</table>

Action prediction accuracy on IkeaDB dataset.
Results: Forecasting Length

Action forecasting accuracies of different lengths, from 0 to 60 timesteps.
Results: Pose

Forecasting results (blue) and ground truth (red)
### Results: Joint Learning

<table>
<thead>
<tr>
<th>Methods</th>
<th>$\text{Acc}_{\text{avg}}$ (%)</th>
<th>$F1$</th>
<th>$E_{\text{pose}}$ (e-2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNN-SW</td>
<td>-</td>
<td>0.60</td>
<td>-</td>
</tr>
<tr>
<td>JCR-RNN</td>
<td>-</td>
<td>0.65</td>
<td>-</td>
</tr>
<tr>
<td>Zero-velocity</td>
<td>87.72</td>
<td>0.67</td>
<td>4.48</td>
</tr>
<tr>
<td>Ours (joint learning)</td>
<td>88.11</td>
<td>0.68</td>
<td>3.84</td>
</tr>
<tr>
<td>Ours (action only)</td>
<td>87.94</td>
<td>0.66</td>
<td>-</td>
</tr>
<tr>
<td>Ours (pose only)</td>
<td>-</td>
<td>-</td>
<td>4.29</td>
</tr>
</tbody>
</table>

Action and pose forecasting evaluation on OAD dataset.