Structured Prediction Helps 3D Human Motion Modelling

-ICCV 2019
-ETH Zürich

June 21, 2020
Chuan Guo
Preliminary

Method

Datasets and Models

Evaluation on Human3.6M

Evaluation on AMASS
Motivation
Contribution

● Main contribution:
  ○ Novel structured prediction layer which incorporate skeleton hierarchy.
  ○ This prediction layer is agnostic to the underlying network.

● Others:
  ○ Evaluations on Human3.6 and AMASS datasets.
Preliminary

Method

Datasets and Models

Evaluation on Human3.6M

Evaluation on AMASS
Structured Prediction Layer

Figure 2: **SPL overview.** Given the context $h_t$ of past frames, joint predictions $\hat{x}^{(k)}_t$ are made hierarchically by following the kinematic chain defined by the underlying skeleton. Only a subset of joints is visualized for clarity.

- $x_t$ is pose vector at $t$ step.
- $K$ is the number of joints.

$$p_\theta(x_t) = \prod_{k=1}^{K} p_\theta(x^{(k)}_t | \text{parent}(x^{(k)}_t), h_t)$$

$$p_\theta(X) = \prod_{t=1}^{T} \prod_{k=1}^{K} p_\theta(x^{(k)}_t | \text{parent}(x^{(k)}_t), h_t)$$
Structured Prediction Layer
Per joint Loss

\[ \mathcal{L}(X, \hat{X}) = \frac{1}{T \cdot N} \sum_{t=1}^{T} f(x_t, \hat{x}_t) \]

Here, \( k \)th joint

\[ \mathcal{L}(X, \hat{X}) = \sum_{t=1}^{T} \sum_{k=1}^{K} f(x_t^{(k)}, \hat{x}_t^{(k)}) \]

Typically, the loss is calculated on pose vector space.

Here, loss is calculated for each joint first, and then summed up for the entire motion.
Preliminary
Method

Datasets and Models

Evaluation on Human3.6M

Evaluation on AMASS
Experiment(dataset)

Input sequences are 2 seconds (120 frames), targets are 400ms (24 frames)

- **Human3.6**
  - 632,894 frames
  - 120 test samples across 15 categories
  - 21 joints

- **AMASS**
  - 9,084,918 frames
  - 3,304 test samples
  - 15 joints
Models

- **Seq2seq**: input poses are represented as **axis angle** (exponential map);

- **QuarterNet**: inputs are **quaternion representation**.

- **RNN**: inputs are **rotation matrices**.
  - Single layer RNN network.
Preliminary

Method

Datasets and Models

Evaluation on Human3.6M

Evaluation on AMASS
## Evaluation on Human3.6M

<table>
<thead>
<tr>
<th>millisecond</th>
<th>Walking</th>
<th></th>
<th>Eating</th>
<th></th>
<th>Smoking</th>
<th></th>
<th>Discussion</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>80</td>
<td>160</td>
<td>320</td>
<td>400</td>
<td>80</td>
<td>160</td>
<td>320</td>
<td>400</td>
</tr>
<tr>
<td>LSTM-3LR [7]</td>
<td>0.77</td>
<td>1.00</td>
<td>1.29</td>
<td>1.47</td>
<td>0.89</td>
<td>1.09</td>
<td>1.35</td>
<td>1.46</td>
</tr>
<tr>
<td>SRNN [14]</td>
<td>0.81</td>
<td>0.94</td>
<td>1.16</td>
<td>1.30</td>
<td>0.97</td>
<td>1.14</td>
<td>1.35</td>
<td>1.46</td>
</tr>
<tr>
<td>Zero-Velocity [20]</td>
<td>0.39</td>
<td>0.68</td>
<td>0.99</td>
<td>1.15</td>
<td>0.27</td>
<td>0.48</td>
<td>0.73</td>
<td>0.86</td>
</tr>
<tr>
<td>AGED [33]</td>
<td>0.22</td>
<td>0.36</td>
<td>0.55</td>
<td>0.67</td>
<td>0.17</td>
<td>0.28</td>
<td>0.51</td>
<td>0.64</td>
</tr>
<tr>
<td>Seq2seq-sampling-sup [20]</td>
<td>0.28</td>
<td>0.49</td>
<td>0.72</td>
<td>0.81</td>
<td>0.23</td>
<td>0.39</td>
<td>0.62</td>
<td>0.76</td>
</tr>
<tr>
<td>Seq2seq-sampling-sup-SPL</td>
<td>0.23</td>
<td>0.37</td>
<td>0.53</td>
<td>0.61</td>
<td>0.20</td>
<td>0.32</td>
<td>0.52</td>
<td>0.67</td>
</tr>
<tr>
<td>Seq2seq-sampling [20]</td>
<td>0.27</td>
<td>0.47</td>
<td>0.70</td>
<td>0.78</td>
<td>0.25</td>
<td>0.43</td>
<td>0.71</td>
<td>0.87</td>
</tr>
<tr>
<td>Seq2seq-sampling-SPL</td>
<td>0.23</td>
<td>0.38</td>
<td>0.58</td>
<td>0.67</td>
<td>0.20</td>
<td>0.32</td>
<td>0.52</td>
<td>0.66</td>
</tr>
<tr>
<td>QuaterNet [25]</td>
<td><strong>0.21</strong></td>
<td><strong>0.34</strong></td>
<td>0.56</td>
<td>0.62</td>
<td>0.20</td>
<td>0.35</td>
<td>0.58</td>
<td>0.70</td>
</tr>
<tr>
<td>QuaterNet-SPL</td>
<td>0.22</td>
<td>0.35</td>
<td>0.54</td>
<td><strong>0.61</strong></td>
<td>0.20</td>
<td>0.33</td>
<td>0.55</td>
<td><strong>0.68</strong></td>
</tr>
<tr>
<td>RNN</td>
<td>0.30</td>
<td>0.48</td>
<td>0.78</td>
<td>0.89</td>
<td>0.23</td>
<td>0.36</td>
<td>0.57</td>
<td>0.72</td>
</tr>
<tr>
<td>RNN-SPL</td>
<td>0.26</td>
<td>0.40</td>
<td>0.67</td>
<td>0.78</td>
<td>0.21</td>
<td>0.34</td>
<td>0.55</td>
<td>0.69</td>
</tr>
</tbody>
</table>

Euler angle metric
Preliminary

Method

Datasets and Models

Evaluation on Human3.6M

Evaluation on AMASS
Evaluation on AMASS

Report accumulated error until time step $t$, instead of error at time step $t$.

- **Joint Angle Difference**: error of rotation matrices
- **Positional error**: error of 3D joint positions
- **PCK**: percentage of predicted joints lying within a spherical threshold $\rho$ around the target joint position
<table>
<thead>
<tr>
<th>milliseconds</th>
<th>Euler</th>
<th>Joint Angle</th>
<th>Positional</th>
<th>PCK (AUC)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100</td>
<td>200</td>
<td>300</td>
<td>400</td>
</tr>
<tr>
<td>Zero-Velocity [20]</td>
<td>1.91</td>
<td>5.93</td>
<td>11.36</td>
<td>17.78</td>
</tr>
<tr>
<td>Seq2seq [20]*</td>
<td>1.46</td>
<td>5.28</td>
<td>11.46</td>
<td>19.78</td>
</tr>
<tr>
<td>Seq2seq-SPL</td>
<td>1.57</td>
<td>5.00</td>
<td>10.01</td>
<td>16.43</td>
</tr>
<tr>
<td>Seq2seq-sampling [20]*</td>
<td>1.71</td>
<td>5.15</td>
<td>9.71</td>
<td>15.15</td>
</tr>
<tr>
<td>Seq2seq-sampling-SPL</td>
<td>1.71</td>
<td>5.13</td>
<td>9.60</td>
<td>14.86</td>
</tr>
<tr>
<td>Seq2seq-dropout</td>
<td>1.26</td>
<td>4.41</td>
<td>9.24</td>
<td>15.46</td>
</tr>
<tr>
<td>Seq2seq-dropout-SPL</td>
<td>1.26</td>
<td>4.26</td>
<td>8.67</td>
<td>14.23</td>
</tr>
<tr>
<td>QuaterNet [25]*</td>
<td>1.49</td>
<td>4.70</td>
<td>9.16</td>
<td>14.54</td>
</tr>
<tr>
<td>QuaterNet-SPL</td>
<td>1.34</td>
<td>4.25</td>
<td>8.39</td>
<td>13.43</td>
</tr>
<tr>
<td>RNN</td>
<td>1.69</td>
<td>5.23</td>
<td>10.18</td>
<td>16.29</td>
</tr>
<tr>
<td>RNN-SPL</td>
<td>1.33</td>
<td><strong>4.13</strong></td>
<td><strong>8.03</strong></td>
<td><strong>12.84</strong></td>
</tr>
</tbody>
</table>

Even a single layer RNN could outperform state-of-art methods on the large and diverse dataset.
Thanks!