We present a learning-based method for generating animated 3D pose sequences depicting multiple sequential or superimposed actions provided in long, compositional sentences. We propose a hierarchical two-stream sequential model to explore a finer joint-level mapping between natural language sentences and 3D pose sequences corresponding to the given motion. We evaluate our proposed model on the KIT Motion-Language Dataset containing 3D pose data with human-annotated sentences. We show that our model advances the state-of-the-art on text-based motion synthesis in objective evaluations by a margin of 50%.

**ABSTRACT**

Our method shows more than 50% improvement on the mean Average Positional Error (APE) and the mean Average Variance Error (AVE) of joint positions over the state-of-the-art methods of JL2P and Lin et al. This research is funded by the BMBF grants XAINES (01|W20005) and IMPRESS (01|S20076), EU Horizon 2020 grant Carousel+ (101017779) and an IMPRS-CS Fellowship. Computational resources provided by the BMWI grants 01MK20004D and 01MD19001B.

**APPROACH**

- We introduce a hierarchical joint embedding space that learns the embeddings of pose and language simultaneously.
- We separate our intermediate pose embeddings hierarchically to limb embeddings such that our model learns features from the different components of the body.
- We have a two-stream sequential network to separately learn the upper and the lower body movements and focus on the end joints of the body.
- We use contextualized BERT embeddings with handpicked word feature embeddings to improve text understanding.
- We further use additional loss terms and a pose discriminator to further improve the plausibility of the synthesized motion.

**QUANTITATIVE RESULTS**

Our method shows more than 50% improvement on the mean Average Positional Error (APE) and the mean Average Variance Error (AVE) of joint positions over the state-of-the-art methods of JL2P and Lin et al.

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean APE w/o trajectory</th>
<th>Mean APE with trajectory</th>
<th>Mean AVE w/o trajectory</th>
<th>Mean AVE with trajectory</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>0.53</td>
<td>0.19</td>
<td>1.06</td>
<td>0.38</td>
</tr>
<tr>
<td>JL2P</td>
<td>1.06</td>
<td>1.13</td>
<td>1.92</td>
<td>1.13</td>
</tr>
<tr>
<td>Lin et al.</td>
<td>1.92</td>
<td>2.56</td>
<td>2.96</td>
<td>2.56</td>
</tr>
</tbody>
</table>

The joint embedding space learned by our method can correlate poses and corresponding sentences better than JL2P and Lin et al. as seen from the Content Encoding Error (CEE) and the Style Encoding Error (SEE) metrics.

*Corresponding Author: anindita.ghosh@dfki.de

**CODE AND RESOURCES:**
https://github.com/anindita127/ComplexText2Animation