Text-Based Motion Synthesis with a Hierarchical Two-Stream RNN

Anindita Ghosh*
Saarland Informatics Campus, DFKI
Germany

Noshaba Cheema
Max-Planck Institute for Informatics, DFKI, Saarland Informatics Campus
Germany

Cennet Oguz
DFKI Kaiserslautern
Germany

Christian Theobalt
Max-Planck Institute for Informatics, Saarland Informatics Campus
Germany

Philipp Slusallek
DFKI, Saarland Informatics Campus
Germany

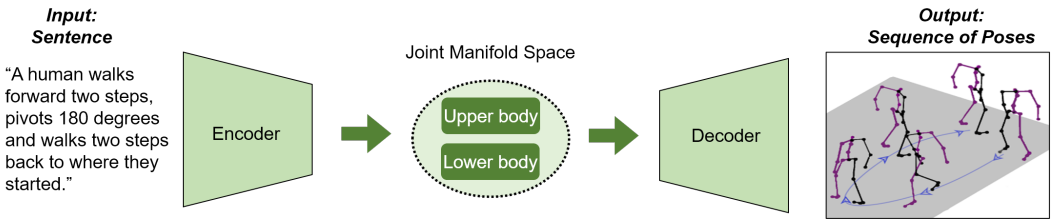

Figure 1: Overview of our proposed method to generate motion from complex natural language sentences. Our model learns a joint embedding for both pose and language, using separate representations for the upper body and the lower body movements.

ABSTRACT

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 the corresponding 3D pose sequences of the motions. We learn two manifold representations of the motion — one each for the upper body and the lower body movements. We evaluate our proposed model on the publicly available KIT Motion-Language Dataset containing 3D pose data with human-annotated sentences. Experimental results show that our model advances the state-of-the-art on text-based motion synthesis in objective evaluations by a margin of 50%.

CCS CONCEPTS

• **Computing methodologies** → **Procedural animation**; *Motion capture*; Natural language processing.

ACM Reference Format:
Anindita Ghosh, Noshaba Cheema, Cennet Oguz, Christian Theobalt, and Philipp Slusallek. 2021. Text-Based Motion Synthesis with a Hierarchical Two-Stream RNN. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Posters (SIGGRAPH '21 Posters), August 09-13, 2021.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3450618.3469163

*email: anindita.ghosh@dfki.de (corresponding author)

SIGGRAPH '21 Posters, August 09-13, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8371-4/21/08.
https://doi.org/10.1145/3450618.3469163

INTRODUCTION

Motion synthesis based on textual descriptions substantially simplifies the task of manually creating realistic animations. It has a rich variety of applications, including language-based task planning for robotics and virtual assistants, designing instructional videos, and visualizing movie scripts [Hanser et al. 2009]. However, mapping natural language text descriptions to 3D pose sequences for human motions is non-trivial. The input sentences, while describing multiple sequential or simultaneous actions, do not correspond to the discrete time steps of the pose sequences to be generated. For example, the input sentence "a person is stretching his arms, taking them down, walking forwards for four steps and raising them again" describes multiple sequential actions, or the sentence "a person s spinning around while walking" describes simultaneous actions, but they do not provide any information on how the pose should look at individual time steps. This necessitates a machine-level understanding of the syntax and the semantics of the text descriptions to generate the desired motions plausibly. Moreover, we need to identify how the different modifiers, such as adverbs and prepositions, impact the output motion. Existing methods for text-to-motion mapping either generate motions that stay in one place [Plappert et al. 2016] or generates simple actions on global trajectories (*e.g.*, walking) [Ahuja and Morency 2019; Lin et al. 2018]. However, these methods fail to translate long-range dependencies and correlations in complex sentences and do not generalize well to complex actions involving synchronized limb movements (*e.g.*,

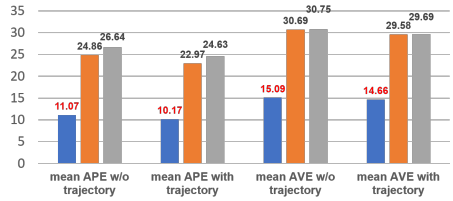

Figure 2: Plots showing the APE (left) and AVE (right) in mm for our method (blue), Lin et al. [Lin et al. 2018] (grey), Ahuja et al. [Ahuja and Morency 2019] (orange). Lower values are better. We observe improvements of above 50% on both metrics for our method.

dancing). We propose a method to handle complex sentences and actions by introducing the following:

- A **hierarchical joint embedding space**, where our model learns 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.
- A **two-stream sequential network** to separately learn the upper and the lower body movements and focus on the end joints of the body (*e.g.*, wrist movements for "playing violin" or footwork for "waltzing").
- **Contextualized BERT embeddings** [Devlin et al. 2018] with handpicked word feature embeddings to improve text understanding.
- A **pose discriminator** with an adversarial loss to further improve the plausibility of the synthesized motions.

Experimental results show that our method outperforms the current benchmarks [Ahuja and Morency 2019; Lin et al. 2018] significantly on both the quantitative error metrics and on qualitative evaluations.

PROPOSED METHOD

We train our model end-to-end with a hierarchical two-stream pose autoencoder, a sentence encoder, and a pose discriminator. Our pose encoder uses five separate layers to focus on the five major parts of the body – left arm, right arm, trunk, left leg, and right leg, and combines them hierarchically to two latent representations for the upper and the lower body poses in the manifold space. We use pre-trained BERT model [Devlin et al. 2018] and LSTM to encode input sentence into similar latent representations as the pose. Our model learns a joint embedding between the natural language and the poses of the upper body and the lower body (Fig. 1). The hierarchical structure of the linear layers in the decoder unit mirrors that of the pose encoder. We add a residual connection between the inputs and the outputs of the individual decoder units such that the decoder learns the velocity of the poses rather than their 3D positions. The GRUs and the hierarchical linear layers in the decoder recurrently output the reconstructed pose for each current frame based on the residual outputs and the latent representations of the previous frames. To train our network, we minimize loss terms describing the error in the pose and the velocity predictions, and the error between the pose and the language embeddings. We also add a pose discriminator with an adversarial loss to further improve the plausibility of the synthesized motions.

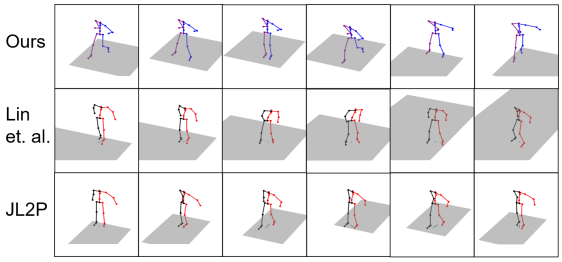

Figure 3: Comparison of consecutive frames of generated animations of our method (top row) with Lin et al. [Lin et al. 2018] (middle row) and JL2P [Ahuja and Morency 2019] (bottom row) for a given sentence. The Waltz dance is prominent in our model. By comparison, in both the benchmarks, the arm movements are missing, and the skeleton tends to slide rather than step.

EXPERIMENTS AND RESULTS

We compare our method with the benchmarks of Lin et al. [Lin et al. 2018] and Joint Language to Pose (JL2P) [Ahuja and Morency 2019]. To quantitatively evaluate the correctness of our motion, we use the Average Position Error (APE), which measures the average positional difference for a joint between the generated poses and the ground-truth pose sequence, and the Average Variance Error (AVE), which measures the difference of variances of individual joints of the generated poses compared to the ground truth poses. Fig. 2 shows more than 50% improvement of our method compared to JL2P and Lin et al. for the mean APE and AVE calculated for all local joint positions with and without the global root trajectory. As an example of the motion quality, Fig. 3 shows that our method generates a prominent waltz dance with synchronized footwork, while both the benchmarks fail.

ACKNOWLEDGMENTS

This research is funded by the BMBF grants XAINES (01|W20005) and IMPRESS (01|S20076), the EU Horizon 2020 grant Carousel+ (101017779), an IMPRS-CS Fellowship. Computational resources were provided by the BMWi grants 01MK20004D and 01MD19001B.

REFERENCES

Chaitanya Ahuja and Louis-Philippe Morency. 2019. Language2Pose: Natural Language Grounded Pose Forecasting. In *2019 International Conference on 3D Vision (3DV)*. 719–728. https://doi.org/10.1109/3DV.2019.00084

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

Eva Hanser, Paul Mc Kevitt, Tom Lunney, and Joan Condell. 2009. Scenemaker: Intelligent multimodal visualisation of natural language scripts. In *Irish Conference on Artificial Intelligence and Cognitive Science*. Springer, 144–153.

Angela S Lin, Lemeng Wu, Rodolfo Corona, Kevin Tai, Qixing Huang, and Raymond J Mooney. 2018. Generating animated videos of human activities from natural language descriptions. *Visually Grounded Interaction and Language Workshop, NeurIPS* (2018), 2.

Matthias Plappert, Christian Mandery, and Tamim Asfour. 2016. The KIT Motion-Language Dataset. *Big Data* 4, 4 (dec 2016), 236–252. https://doi.org/10.1089/big.2016.0028