Topic B - 3D Human Motion Generation

November 2024

1 Introduction

The development of 3D human motion generation is an important advancement in fields such as film, video games, and virtual reality. Traditionally, this process demands some expertise, often necessitating the manual work of a human animator. The progress of motion generation models could assist them during this process, enabling beginner animators to generate realistic motion and experts to work more efficiently. This proposal focuses on enhancing motion generation models using data augmentation, particularly through text augmentation. By leveraging these techniques, we aim to produce models capable of synthesizing lifelike human movements.

2 Plan of works

We will start with the literature review. We study key references related to text-based motion retrieval [1], human motion diffusion models [4], and timeline control for text-driven motion generation [3]. In addition to them, I studied Text-to-motion retrieval using contrastive 3D human motion synthesis [2].

The datasets used in this project are constructed using a shared motion data source, AMASS, and annotated differently in HumanML3D, KIT-ML, and Babel. A single motion in AMASS can appear in one or more of these datasets but will have different text annotations depending on the dataset.

- Motion data: We will download the unannotated AMASS dataset using instructions from the STMC repository.
- Text annotations: The textual annotations for HumanML3D, KIT-ML, and BABEL will be retrieved from the TMR++ repository, which includes the augmented annotations.
- Text embeddings: For compatibility with the MDM-SMPL model, we will generate the text embeddings corresponding to the annotations following instructions from the STMC repository.

Afterwards, we will establish a baseline against which improvements can be measured. To do so, we will evaluate the current performance of pretrained MDM-SMPL models from STMC on the HumanML3D, KIT-ML and Babel. Then We will realize three trainings : (1) on KIT-ML, then (2) on HumanML3D and KIT-ML and finally (3) on HumanML3D, KIT-ML and Babel. We will test these three trainings on HumanML3D, KIT-ML and Babel datasets to build another baseline.

Then we will explore the impact of text data augmentation on motion generation. This involves training the models on HumanML3D, KIT-ML, Babel with text augmentation (annotations from TMR++) and test on HumanML3D, KIT-ML and Babel datasets.

An extension to the existing method will be proposed by implementing techniques such as dropping or repeating frames, varying the frame rate, adding noise to motion data and encoding-decoding motion data with added noise in the bottleneck.

The project will include a comprehensive experimental evaluation, both quantitatively using standard metrics, and qualitatively by visualizing the motions synthesized from various textual prompts, with comparisons to ground-truth motions and which types of texts and motions work best. This dual approach allows a deep understanding of where the model achieves good performance and where improvements are necessary.

3 Expected difficulties

Setting up advanced models and ensuring proper dataset preparation can be technically challenging, especially with complex integrations like text augmentation. Additionally, the project may face computational resource constraints since we have limited access to GPU. Also, it could be challenging to achieve robust generalization of the model.

References

[1] Léore Bensabath, Mathis Petrovich, and Gul Varol. A cross-dataset study for text-based 3d human motion retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1932–1940, 2024.

[2] Mathis Petrovich, Michael J Black, and Gül Varol. Tmr: Text-to-motion retrieval using contrastive 3d human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9488–9497, 2023.

[3] Mathis Petrovich, Or Litany, Umar Iqbal, Michael J Black, Gul Varol, Xue Bin Peng, and Davis Rempe. Multi-track timeline control for text-driven 3d human motion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1911–1921, 2024.

[4] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023.