Multimodal 3D Human Body Generation and Interaction

1 Research topic

The research topic of this thesis proposal falls in the domain of **generative AI**. In particular, we aim to develop a mathematical framework for **multimodal generation of 3D human bodies and their reciprocal interaction or interaction with objects**. In doing this, we will touch upon some of the most challenging problems in different research fields such as **Computer Vision** and **Graphics**, where generative models are very crucial. The research topic itself is very timely in terms of the need and applicability of the targeted systems. This research also has the ambition to advance fundamental tools that are not only of high relevance in terms of intellectual merit but also of broad impact.

The methods and tools developed in the thesis will find potential application in several contexts such as the generation of synthetic data for augmenting the training capabilities of machine learning solutions, for example in robot and human-robot interaction; for generating avatars in the movie industry; for augmented and virtual reality applications, for simulation, metaverse, etc.

2 Objective

AI approaches have shown very promising results in generating facial expressions and movements in different interacting contexts. Otberdout et al. in [5] and [4], for instance, proposed a new approach for generating 3D facial expressions. However, these works can only generate a sequence of faces for the corresponding subject, and not the facial reactions of the interlocutor. Several scientific challenges need thus to be addressed to solve the generation of human-appropriate non-verbal reactions in various interaction scenarios. The goal of 3D human body generation is to create virtual humans that closely resemble real individuals in terms of appearance, pose, and movement. This has applications in various fields, including Computer Graphics (for creating realistic avatars in movies and video games), Virtual Reality, Augmented Reality, and Human-Computer Interaction. Chopin et al. [2] proposed InterFormer, a new Transformer [6] for generating human reactions. Although the results obtained are very encouraging, this work has scientific and technical limitations that we will address in this thesis. From a scientific point of view, this work has several limitations because the method is not capable of generating a large diversity of

movements, which limits its applications. However, with more diversity, evaluating the results becomes difficult. Finally, when generating longer sequences, the movements lack coherence with respect to the action's movement. The recent results obtained by [1] show that modeling interactions with graphs combined with diffusion models generates coherent sequences of human movements. These preliminary results demonstrate the value of this approach for modeling interactions and non-verbal reactions. Ghosh et al. [3] introduced ReMoS, a denoising diffusion-based probabilistic model for reactive motion synthesis that explores two-persons interaction. Given the motion of one person, the reactive motion of the second person is synthesized to complete the interactions between the two. However, all these works do not explore human-object interactions.

In this thesis proposal, we target *multimodal* 3D human body generation from text and/or audio inputs as well as the generation of 3D humans that interact with other humans or objects. We aim to design a generative solution, where textual description can be used to describe the movement and the interaction accounting for spatial/temporal behavior. Large-language models will be considered as a way to pass from a high-level textual description to a finer and more detailed one, thus also injecting additional knowledge in the generative process. Audio will be also regarded as a speech track, and so an alternative way to obtain a textual description for 3D human generation. In addition, in our view it could be also used to represent a way to generate synchronized audio-body movement for synthesizing animations like for a dancing action. This poses several challenges in the way different modalities can be combined together, obtaining natural and smooth generation and dynamic human-human and human-object interaction.

Addressing these challenges is essential for creating realistic and contextually relevant animations. In particular, we will aim to address the following key points:

- **Capturing subtle human body motion**: Human body motion is complex, nuanced, and influenced by various factors such as emotions, intentions, contextual cues and constraints set by the environment and objects therein;
- **Text to human-human and human-object generation**: text can be inherently ambiguous, and the same textual description may be interpreted differently by different individuals. Resolving ambiguity in the interpretation of the text to generate accurate and intended motions and interactions is a significant challenge;
- **Audio to 3D human motion synchronization**: Aligning audio cues with corresponding 3D human motions is a fundamental challenge;
- **Expressiveness and emotion**: Capturing the emotional content of audio, and translating it into expressive human motions is complex;
- **3D human-object interaction generation**: Existing works generally focus on generating dynamic human beings, without taking their environ-

ment into account. In this thesis, we will consider the interaction between humans and the environment.

3 International collaboration

This thesis is a Cotutelle thesis between the teams 3D-SAM of CRIStAL and the Media Integration and Communication Center (MICC) of the University of Florence. The 3D-SAM team from CRIStAL and the MICC have established a continuous, solid, and fruitful collaboration over more than 10 years. They have published more than 20 papers in top journals and conferences (TPAMI, IEEE TAC, CVPR, IEEE FG). Since 2022, the two groups are funded by a CNRS International Research Project (IRP) GeoGen3DHuman (2022-2027). The candidate will be involved in this International project.

4 Required skills for the PhD candidate

The following background knowledge and skills are required for the PhD candidates:

- The applicant should had conducted Master or engineering studies in relevant fields (artificial intelligence, data science, computer vision, mathematics).
- Strong algorithm and programming skills (PyTorch/Keras/Python/etc.).
- Very strong knowledge in computer vision and deep learning techniques (CNN, GAN, Diffusion models, etc.).
- Knowledge on 3D data processing is appreciated.
- Fluency in written and spoken English. Fluency in French/Italian is not required.
- Relational working qualities.

5 Information and Contacts

- CRIStAL
  - Mohamed Daoudi, Professor (mohamed.daoudi@imt-nord-europe.fr)
  - Deise Santana Maia, Assistant Professor (deise.santanamaia@univ-lille.fr)
- Media Integration and Communication Center (MICC)
  - Pietro Pala, Professor (pietro.pala@unifi.it)
  - Stefano Berretti, Associate Professor (stefano.berretti@unifi.it)

Important dates:

- Application deadline, May 1st, 2024.
- PhD starting date, October 2024.

6 Application

The application process involves sending the following information, to mohamed.daoudi@imt-nord-europe.fr, deise.santanamaia@univ-lille.fr, pietro.pala@unifi.it, and stefano.berretti@unifi.it, subject [Multimodal3DHuman]

- A detailed Curriculum vitae. Ensure that your CV highlights your relevant experience, skills, and accomplishments,
- The last three years of University transcripts (bachelor, master) and diploma,
- Motivation letter,
- Two Recommendation letters.

Once you have submitted your application, you may be contacted for a remote interview.

References

[1] Baptiste Chopin, Hao Tang, and Mohamed Daoudi. Bipartite graph diffusion model for human interaction generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5333–5342, January 2024.

[2] Baptiste Chopin, Hao Tang, Naima Otberdout, Mohamed Daoudi, and Nicu Sebe. Interaction transformer for human reaction generation. *IEEE Transactions on Multimedia*, 25:8842–8854, 2023.

[3] Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. Remos: Reactive 3d motion synthesis for two-person interactions, 2023.

[4] Naima Otberdout, Claudio Ferrari, Mohamed Daoudi, Stefano Berretti, and Alberto Del Bimbo. Generating multiple 4d expression transitions by learning face landmark trajectories. *IEEE Transactions on Affective Computing*, pages 1–12, 2023.

[5] Naima Otberdout, Claudio Ferrari, Mohamed Daoudi, Stefano Berretti, and Alberto Del Bimbo. Sparse to dense dynamic 3d facial expression generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20385–20394, June 2022.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017.