Literature Review of Modern Motion and Dance Synthesis Techniques and Their Benefits and Flaws

Jiazhi Zhou

April 18, 2024

Abstract

This literature review contributes to the research towards better embodiment experience in VR technology. The results of this review will allow the research team to deploy the appropriate model for the given purpose of the experiments and observe the reactions of users to the model.

1 Introduction

TODO: write

2 Literature Review

A literature review is conducted on the domain of human-to-robot interaction and AI generative dances. The goal of the review is to gain an initial understanding of the work that is being done in these domains, as well as providing insights into what could have been done better, existing models that work well, and how we can model human-to-robot dance interactions well.

2.1 Interaction and Turn-taking

Turn-taking is an important concept to consider when thinking about interactions between human and robots (AI). When having conversations, it is second nature for us, when a sentence is about to end, and when the other person, who is currently the "follower" or who is listening to the "leader", would have a chance to speak. But dance turn-taking is not the same as speech turn-taking [7], as in the setting of improvised dance, turn-taking happen naturally, and the leader and follower concept are less apparent, since both parties are dancing at the same time. Understanding the dynamics of leader and follower, how turn-taking occurs in dance, and how the user performs while being a leader and a follower is critical for human and computer to be able to interact naturally.

Turn-taking in speech is a good topic to study and some important aspects of turn-taking could be transferred from speech turn-taking to dance turn-taking. Thomaz and Chao conducted an experiment to learn about turn-taking in human-robot interaction [5]. The user study is setup by having a human and a "robot" play a game of simon says, the robot, being teleoperated by a human. The researchers judged each potential turn-ending indicator against the time it took for the human to react to the sentence and start their action. All indicators that had a "negative" reaction time were deemed impossible. This resulted in the concept of Minimum Necessary Information (MNI) as the primary indicator in this setting, as humans can start doing an action after all the necessary words of the sentence are said and the rest is redundant. This however is not in an improvisational setting, as all actions and all cues are pre-set, and it is easy to predict the following words.

Skantze did a analysis for conversational turn-taking, which is a under a more improvisational setting [4]. The concept of prediction and reaction are further explored. Prediction was said to play an important factor in human to human speech interactions, which is what makes it so smooth, but provid-

ing turn-yielding cues are also an important step in the turn-taking process. The turn-yielding cues are a way for the leader to yield their current turn to the follower, and the effects of using the cue is said to be additive, so more turn-yielding cues result in a better understanding by the follower. But prediction is also an important factor, just like it was discovered in [5]. The prediction for the experiment ran by Thomaz and Chao are the MNI point, where the human can predict the rest of the words, as the phrases are pre-set.

When considering how turn-taking should occur, it is critical to observe the difference of turn-taking done by professional to that of the average person, as different frameworks might have to adopted for the purpose of a public installation vs for co-creative dance agent. The difference how experienced dancers and non-performers communicate and turn-take is studied by Evola et al. to gain understanding of how experts are able to come to an agreement on the leadership state without much inter-subjectivity. A improvisational performance is used as the framework for the user study, where users in a group of 6 can take turns to construct an art piece. The turn-taking sections of the performances are analyzed. The expert performers did not performing any "communicative movement", but rather using observations from their parafoveal and peripheral vision to take cues, while the non-performers were seen exchanging gazes to communicate. The ability to use context clues to understand turn-taking cues seem to align with the research into speech turn-taking, which explains how the expert performers' turn-taking and human-to-human speech turn-taking can be performed with so much fluidity. But when less experienced people that do not have a shared "language" or cues try to perform turn-taking, they perform less naturally.

Turn-yielding cues, as well as turn-taking cues are also a concept in dance turn-taking explored by Winston and Magerko [7]. These cues are explored via implementing a turn-taking system on the LuminAI framework, to create a new version that they named, TT-VAI. The base version of LuminAI and TT-VAI are studied in a user study. Certain motion cues are used by TT-VAI as a turn-taking cue, including energy, tempo, and size. Out of the two versions of the model, the turn taking model was disliked by a few users, who preferred less "back-and-forth" and preferred more natural interactions and felt more inspired. 2 users preferred the turn-taking one as it was providing more than just copying. Mimicry by the agent, however, showed positive feedback from several participants of the user study, since the agent is deemed "more responsive to my movement." The research highlighted interesting ideas in how a sense of leadership state effect's the user's perception of the model, and how mimicry or being a follower by the model can provide benefits for improved user perception. Tuning the agents turn-taking decision making process, and potentially biasing it towards humans leading could be done to improve the user experience in future works.

TODO: write about B Wallace's work The concept of breaking shared images from Wallace et al.'s work, can fit into the turn-taking paradigm where breaking of shared image is a way that the partner has used to take the leadership and shift the direction of the dance.

The theme of AI bringing more to the table is explored in both Wallace et al., and Winston and Magerko's work [6, 7]. The dance professionals from [6] expressed that they would expect their dance partner to be able to shift the tone of the improvisation to generate more inspiration and keep the improvisation going. This means bringing something different, like doing the opposite of what they are doing. Similarly, [7], although only 2 of the users expressed this, it is still a point made, during interactions with TT-VAI, that, they liked it when the AI was able to do more than just mimic. And although mimicry had a positive reaction by the non-dancers from [7], it was observed from the experiment in [6], that the dance professionals rarely mimicked their partner but sometimes copy things like the trajectory of movement or mirroring their use of space. This means something that the public could be into, might be less inspiring for professional dancers. This could mean that even with a user led interaction, the AI can take the role of the leader when the user seem disengaged and running out of ideas, then taking the dance in a different direction to potentially generate some more interest and inspiration.

Another new concept for turn-taking that is particular to dance, is the perception of leadership state. As in speech, it is obvious who is leading and who is following, but in fluid interactions like dancing, the leadership state has to be inferred by both parties. This could result in a conflict of state, where multiple leaders exist or no-leaders exist at once. The work by Winston and Magerko only considered the AI agent's perception of who the leader and follower is, and did not take into account the user's perception. This should also be explored further as a way to improve the user's experience.

The turn-taking model of expert dancers can potentially be modeled to allow a co-creative dance agent to understand turn-taking cues and perform seemlessly with dancer professionals. The turn-taking for non-dancers, however, is more complicated, as chances for the non-dancers to learn the cues and be able to integrate that into their mental model is likely impossible. However, as Winston and Magerko suggested, casual users are likely more into user led interactions, or at least biased towards that. The difference of expert dancers and casual users interacting with a turn-taking model like TT-VAI should be investigated to gain insights into if turn-taking should be considered at all for a dance robot or should it be full or biased towards user led interaction.

2.2 Dance AI and deep learning

Machine learning and deep learning has been a popular topic in recent years. Showing incredible results for text generation, image generation and much more. Using deep learning techniques to train machine learning models could enable the generation of realistic responses to user's full body inputs in a VR or public installation setting. This can not only prompt the users to explore different movements in order to have a better embodied experience, not also prompt user interactions with the installation, or become a potent tool for co-creative purposes [**?**]. Different techniques and models will be reviewed and examined on their abilities to generate realistic, diverse and real-time dances for the purpose of an interactive AI agent.

Alemi et al. explored the idea of an interactive AI agent [1]. Alemi et al. compared the Factored Conditional Restricted Boltzman Machine (FCRBM) and a Long Short Term Memory (LSTM) network. And at the time, there was not a large public annotated dance dataset available like AIST++ [2], so Alemi et al. had to record their own dance data which only consisted of 4 dance performances and a total of 23 minutes of dance and audio data.

Bailando++ is neural netowrk model that generates dances based off of the previously generated dance sequences [3]. The Bailando++ model is a VQVAE and a Generative Pre-trained Transformer (GPT) that generates dances from a previous dance sequence and dances in sync to the music. The model is trained to dance to the music through the "Actor Critic" learning stage which leverages reinforcement learning and uses beat-alignment as part of the reward function. This model was able to achieve top-of-the-line results in motion quality, as well as motion diversity compared to other popular dance AI models at the time, including DanceNet, DanceRevolution, FACT and Li et al. Bailando++ also preformed well in the user study where users are shown 60 pairs of dances by different models and voted on "which one is dancing better to the music", where it was able to achieve at least 88% win rate against all of the models. Although Bailando++ focused heavily on the ability to dance to the music, it is likely that for the purpose of our experiments, the ability to dance to the music is not as important, as care more about the ability to prompt interaction. But the technique of using actor-critic learning can be used in our own model.

Dance with you (DanY) [8] is a neural network model that generates dances for a partner dancer for a lead dancer. The model uses a three stage network that also leverages a VQVAE for encoding and decoding, and U-Net Models. VQVAE is an auto encoder network which can turn complicated dance data sequences into quantized codes from a finite code book that is learned through the training of the VQVAE, and the U-Net takes noised data and turns them into dances features in the code book, which turns random gaussian noise into realistic dances. The difference of the DanY model is that not only does it

generate from the condition of audio data like many other models, but it also generates based on the condition of the lead dancer's dance sequence. This is important for us since we want to deploy an interactive AI agent which dances in accordance to the lead dancer, who, in this case, is the user. The quantitative results from this Their proposed AIST-M dataset is a dance dataset that contains Lead-Partner dancer pair annotation great for training models to generate partner dances from a lead dance sequence. The techniques they used followed that of the creation of the AIST++ dataset [2] including tracking, SMPL mesh fitting, and optimization for filtering out undesirable frames to ensure the quality of the dance data. The proposed AIST-M dataset will be incredibly useful for our own models' training and analysis.

References

[1] Omid Alemi, Jules Françoise, and Philippe Pasquier. Groovenet: Real-time music-driven dance movement generation using artificial neural networks. 4 2017.

[2] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. 1 2021.

[3] Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando++: 3d dance gpt with choreographic memory. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:14192–14207, 2023.

[4] Gabriel Skantze. Turn-taking in conversational systems and human-robot interaction: A review, 5 2021.

[5] Andrea L Thomaz and Crystal Chao. Turn taking based on information flow for fluent human-robot interaction. 2011.

[6] Benedikte Wallace, Clarice Hilton, Kristian Nymoen, Jim Torresen, Charles Patrick Martin, and Rebecca Fiebrink. Embodying an interactive ai for dance through movement ideation. pages 454–464. Association for Computing Machinery, 2023.

[7] Lauren Winston and Brian Magerko. Turn-taking with improvisational co-creative agents, 2017.

[8] Siyue Yao, Mingjie Sun, Bingliang Li, Fengyu Yang, Junle Wang, and Ruimao Zhang. Dance with you: The diversity controllable dancer generation via diffusion models. pages 8504–8514. Association for Computing Machinery, Inc, 10 2023.