Literature Review

Jiazhi Zhou

April 22, 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 Existing work

TODO: write

2.1 Human-Co-Creative-Agent Interaction

Interaction design in Human-Robot co-creative systems are important. Studies have shown that a system with good interaction design, like providing feedback, allowing for multiple people to interact, etc. have proven to improve people's perception and attitude towards to the AI, and be more inspired, and lead to facilitating positive social interactions [5, 6]. Kantosalo et al. said that: interaction design should be ground zero for designing co-creative systems [3]. Despite this, many AI research failed to consider this, and focus on the model but neglect the interaction [7]. So to ensure our agent will have high quality interactions with the users, we gather important interaction design requirements from all the critical stakeholders, including professional dancers, and the general public alike. This will first be done, through an literature review and analysis of critical modern co-creative AI systems. And a dive into the interaction design and analysis space will be done, in order to facilitate a more systemic way to design and study these interactions.

2.1.1 Requirements and Expectations on Co-Creative Interaction Systems

Wallace et al. performed a workshop studying the expectations from professional dancers for an AI dance partner [10]. The workshop introduced to the dancers on a high level, how a dance generation AI would work. The dancers are then put into pairs, where one of the dancer from the pair would explore the role of an AI dancer, and the other one would explore the role of the user of the AI. The workshop ended with a qualitative feedback session which resulted in important insights into co-creative process for dancers. One important theme that a lot of the dancers mentioned in their feedback was "shared images", more specifically, having a shared language and understanding of their work, and then, breaking of the image, allowing them to continue the creative process exploring other directions, when the current creative direction runs dry. Another theme was that human dancers would eventually get tired, or experience physical constraints to their body, but they thought AI dancers can explore impossible positions which can act as inspiration rather than a negative. The final theme was that dancers have the intuition to perdict what their partner will do, resulting in a more fluid interaction process and allowing them to not look at each other all the time. This also comes from having a shared image or understanding of the creative process. These themes should be taken into consideration for how a dance generation AI should behave from the perspective of professional dancers, however, this could differ for non-dancers.

LuminAI is a good example of existing modern co-creative dance agent. In a 2019 article, the creators of LuminAI, aka Long et al., analyzed many existing co-creative agents in public installations (including LuminAI), and identified interaction design principles that creates a better interaction experience. There are many design principles that this article identified, including technology design and research design, we are mostly interested in the design principles related to interaction design. The first interaction design principle is to have multiple entry points for interaction. What this means for our research, is to consider interaction of both non-dancers as well as dancers to have a good experience, which should allow for a more inclusive experience. The second point, is to allow for group interaction, and also try to, through the installation, facilitate human collaborations. This could mean, for the non-dancers, bring people to dance together on the floor, and for dancers, allow multiple dancers to dance together. The ability for the model to allow for both experienced dancers and non-dancers raise a interesting challenge. As we have identified many differences in the preferred interaction styles of the general public compared to professional dancers. But designing with the use of both groups in mind can help bridge the gap and allow for a general design. Allow for multiple dancers to interact becomes a challenge to the AI model, but there can be ways to average the data before or after the information gets passed to the model for inference, that can be used to avoid increased training complexity.

Winston and Magerko explored the interactions of non-dancers with the aforementioned interactive public installation: LuminAI, and designed a new interaction framework for LuminAI and named it TT-VAI [11]. TT-VAI differed from the base version, as it not only follows the user but sometimes would try to lead the dance. The AI would uses metrics such as energy, temp, and size, as turn-taking cues, to know when to take the lead, and when to give it back to the user. Through observations, Winston and Magerko identified common turn-taking cues that dancers use, when performing lead-partner dances. This included gaze, haptic feedback (in the form of applying pressure to the follower's arm), and energy. Gaze and haptic are left out of the decision making model of TT-VAI as those two are not good communication channels for a projected agent. The base version of LuminAI and TT-VAI are studied in a user study. 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. Two users preferred the turn-taking model as it was providing more than just mimicking. However, mimicry 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. It also highlighted issues with this interaction framework, like the existence of intersubjectivity for who the leader is.

Turn-taking in improvisational settings are studied by Evola et al. [2], where clear distinctions between how turn-taking happens for seasoned dancers, and non-dancers are observed. An improvisational performance is used as the context for the user study, where users in a group of 6 can take turns to construct an art piece from an assortment of items placed on a table in the centre of the stage. The turn-taking sections of the performances are picked out and analyzed. One major different identified, between expert performers and the non-dancers, were 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. This revealed that expert dancers can be way more intuitive during an improvisation which leads to a fluid and seemless performance, while non-dancers often hesitate, and want more explicit communication and turn-taking cues. In trying to explain the non-dancer's hesitence, another idea generated, was the non-dancers wanted approval and to please the choreographer, which is why there was an increased amount of hesitation. This is possibly resulted from the context of the experiment, where the turn-taking is in a ”performance” which has a certain level of strictness engrained in people’s minds, leading to nerve.

The idea of intuition flow aligns well with what Evola et al. discovered from their experiment with the professional dancers [2, 10]. The idea of intuition was the result of a discussion involving how professional dancers were able to remove intersubjectivity during the turn-taking excersize without using communicative movements. Evola et al. discussed how the dancers perhaps viewed the excersice not as turn-based creation of art, but just the creative process in itself, so there was no thoughts of when should I take a turn, but rather, everyone were on the same page, and were connected, allowing them to form a ”coordinated communication behaviour”. The way Wallace et al. discussed this behaviour was that: dancers do not think, but simply do. The intuition is what tells the dancers what to do during an improvisation. This could mean that we can try to train AI to pick up on the ”coordinated communication behaviour” and act like they understand as well, or use a different interaction framework, and simply have the AI try to contribute to the ”art piece” (the dance), and use a feedback system to know how well it is interacting, and contributing.

It is, however, crucial to think about how the intuition flow comes into play with non-dancers. From Evola et al.’s study, we realized that non-dancers were not using a ”coordinated communication behaviour”. This problem also surfaced from Zamm et al.’s study [13], where users attempt to learn to tap to a rhythm in a pair. The study first had users learn to tap by themselves, then they were put into pairs. The result showed that the users were unable to produce the same result as they did by themselves, in a pair, even when they have had some practice by themselves already. This shows that the ability for non-dancers to grow the intuition to move with any dance partner without thinking, over the course of playing with the installation, or in any short user study settings, is likely impossible. However, the need for a smooth interaction was highlighted from Winston and Magerko’s work, where users that preferred the base version of LuminAI thought it had less ”back and forth”, and preferred a more natural interaction. This means, people do not dislike turn-taking, but simply dislike the unnaturalness to it. This view aligns with the study of conversational turn-taking [9], where traditional conversational turn-taking interaction between human and robots feel awkward because of either pausing too long, or talking over the human because of awkward and non-agreed turn-taking cues. But since non-dancers are unable to turn-take well with others either, since they are unexperienced, perhaps it is good to have turn-taking be an optional feature and simply have the AI try to produce as natural interaction as possible.

While the experts want an improvisational dance to be about the creative process, general public could want different things. From Winston and Magerko’s work, we learn that mimicry was preferred by some of the participants, as they liked it when the agent was responsive to their movements. This is true in some levels for professional dancers, as Wallace et al. observed that the dancer pairs would often indirectly mimic their partner. This means mimicing of the movement trajectory etc. but never mimicking the movement exactly. As Evola et al. and Wallace et al.discused, the process of creation is not viewed by the dancers as granular interactions but the dance as an art piece as a whole. So interaction framework for interacting with a professional dancer should be about the creative process and the ”product” (in a non-physical sense), while interaction with the general public could be more on a granular level, where certain movements or ways of interactions could be more inspiring to facilitate group interaction and get more people to be involved.

Wallace et al. discussed, how the existence of glitches could positively effect the professional dancers, since it gave them new ideas. However, the glitching when shown to the general public, could be an issue as glitchy movements could potentially deter users and make them think the AI is faulty and bad. As the main goal of professionals using the AI would be to get inspired and keep the improvisation going, while the main goal for a public installation would be to get people involved and be inclusive, these differences could mean that only one model can be explored in this research project due to time and resource cost. However, the design principle from [5] also mentioned that public installation AI systems should be designed with modularity and maintainablity in mind, so if we use the principles of modularity, we can potentially design general modules that both of these interaction frameworks will be able to adopt, and then have specialized modules that differ between the professional model and the public installation model. This would also allow the agent to be shown to a wider range of people, allowing for more feedback and expose more people to co-creative AIs.

2.2 Interaction Framework

Rezwana and Maher proposed a novel framework for designing co-creative interaction called the COFI which can help put the requirements gathered from the above section into more concrete terms, as well as confirming some ideas that we had. This framework breaks the co-creative interaction down to low level aspects. This includes: what collaboration style and communication style does the framework have, is the creative process a generative process, or a evaluation process or a definition process, how does the AI contribute to that process, and how much should the AI contribute compared to the human contribution. The exact framework breaks this down in a bit more detail, but as an overview, COFI allows us a way to group different interaction frameworks and compare them to each other. After proposing the COFI framework, Rezwana and Maher identified and evaluated a list of co-creative systems and sorted them using COFI. The results highlighted 3 Human-AI co-creative interaction frameworks. The first being an agent to follows and complys with human contributions, and generate similar contributions. This is a pleasing agent, which can be good for extending human's creativity. The opposite of a pleasing agent is a provoking agent, but both pleasing and provoking agents play important roles in co-creative systems. The second is models that uses spontaneous initiative-taking and creates in parallel with users. This would align more with the improvisational dance agent. However, the main issues with this domain of models is the lack of communication flow from the agent to the user, which can negatively impact the quality of collaboration and user experience. The thrid cluster of models is an advisory and evaluation type of agent, where the agent would provide feedback and can refine the user's deisgn. However, this model lacks the communication channel from user to the agent. Overall, the three main clusters all have their pros and cons, and ensuring good communication between the agent and the user would be crucial for a good user experience and collaboration quality.

2.2.1 Analysis

In the first cluster of models, where the agent would follow and comply with the human contributions could be looked into for a co-creative agent. The difference between how professional wants to interact with the AI agent and how the public would likely want to interact with the AI agent fits nicely into the COFI framework. The COFI framework analysis brought up the ideas of "pleasing agent" and "provoking agent". Where pleasing agent can perform the action of extending one's creativity, like Text-To-Image generation AIs bringing user's idea to life. The place for pleasing agent for professionals in general, however, is uncertain, and a specific user study for this can be conducted for professional dancers in order to learn whether the "pleasing" quality of a co-creative agent could also benefit professional dancers in achieving their creative ideas. Pleasing agents are also identified as being able to follow the user's style and artistic vision. That aspect might fit better into the use of a co-creative agent for improvisational dancers. Provoking agents, like using glitches to inspire dance [10], and "breaking of shared images" seem to fit better for professionals where their goal is to generate new ideas. However, with the non-dancers from Winston and Magerko's work, out of the 6 people who understood which version mimicked more, 2 of them liked TT-VAI (turn-taking version of LuminAI) better, quote: "seemed like it was doing more" and “seemed more ready to throw something into the party.” This demonstrates that a provoking agent might not strictly be applicable for professionals while a pleasing agents also might not be strictly for the general public.

The second cluster of co-creative agent interaction framework has a more aligned theme with our study of improvisational dance agent. The idea of mimic and non-mimic agent is a critical part of this cluster of agents. As discussed previously, mimicry can have its place in both non-dancer and dancer interactions, but no strict mimicry of moves should be used when interacting with a professional dancer. However, a big pitfall with these models, is a lack of Agent-To-Human communication which negatively impacts the interaction. The importance of a two-way communication is studied more concretely by Rezwana and Maher in an earlier article [6], where a user study demonstrated that adding a communication channel from agent to the user can significantly improve user experience and perception to the agent, and improve the users' ability to create. However, communication in an embodied way is a challenging task in itself, since we saw that professional dancers often does not "communicate" but rather use their instincts to work together. This makes creating a meaningful Agent-To-Human feedback channel rather challenging. But using other means of communications that the user could understand better, like visuals, audio, and even text could be attempted in order to study its proficiency and effect on the interaction. This communication channel could also be implemented as a module for ease of swapping in and out, for different use cases of the model.

The third cluster, of an advisory and evaluation agent could be counter intuitive for an embodied interaction. However, the fact the cluster three and cluster two complements each other prompts the ideas of combining two agents into one, where one would simply dance, and another would evaluate the creative process and creative product which would act as a way to give feedback to users during the process. The mode of communication for Agent-To-Human, would still be important to figure out. This also prompts the thinking of where does this agent live, as a visualisation could live right in the 3D space, but a UI could potentially break immersion or even distrupt an professional dancers intuition. The study of effective ways to deliver.

2.3 Interaction Labeling and Analysis

Interaction, especially embodied interactions like dance, is a challenging thing to analyse. But having proficient tools to analyze the data after gathering information will be critical in generating further insights, and comparing different models against each other on meaningful metrics. This prompts a look into recent, and novel methods for measuring, analyzing and labeling co-creative, and embodied interactions.

2.4 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 [10]. 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++ [4], 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 [8]. 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) [12] 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 [4] 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.

3 Discussion

TODO: write

3.1 Interaction Design

TODO: write

3.2 AI Model With Regards to Interaction

TODO: write

3.3 Experiments and Analysis

TODO: write

4 Summary

TODO: write

References

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

[2] Vito Evola, Joanna Skubisz, and Carla Fernandes. The role of eye gaze and body movements in turn-taking during a contemporary dance improvisation.

[3] Anna Kantosalo, Prashanth Thattai Ravikumar, Kazjon Grace, and Tapio Takala. *Modalities, Styles and Strategies: An Interaction Framework for Human–Computer Co-Creativity.* ACC = Association for Computational Creativity, 2020.

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

[5] Duri Long, Mikhail Jacob, and Brian Magerko. Designing co-creative ai for public spaces. pages 271–284. Association for Computing Machinery, 2019.

[6] Jeba Rezwana and Mary Lou Maher. Understanding user perceptions, collaborative experience and user engagement in different human-ai interaction designs for co-creative systems.

[7] Jeba Rezwana and Mary Lou Maher. Designing creative ai partners with cofi: A framework for modeling interaction in human-ai co-creative systems; designing creative ai partners with cofi: A framework for modeling interaction in human-ai co-creative systems. *ACM Trans. Comput.-Hum. Interact*, 30, 2023.

[8] 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.

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

[10] 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.

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

[12] 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.

[13] Anna Zamm, Stefan Debener, and Natalie Sebanz. The spontaneous emergence of rhythmic coordination in turn taking. *Scientific Reports*, 13, 12 2023.