User Motion Accentuation in Social Pointing Scenario

Ruoxi Guo*
University College London

Lisa Izzouzi†
University College London

Anthony Steed‡
University College London

Abstract

Few existing methods produce full-body user motion in virtual environments from only the tracking from a consumer-level head-mounted-display. This preliminary project generates full-body motions from the user's hands and head positions through data-based motion accentuation. The method is evaluated in a simple collaborative scenario with one Pointer, represented by an avatar, pointing at targets while an Observer interprets the Pointer's movements. The Pointer's motion is modified by our motion accentuation algorithm SocialMoves. Comparisons on the Pointer's motion are made between SocialMoves, a system built around Final IK, and a ground truth capture. Our method showed the same level of user experience as the ground truth method.

Index Terms: Human-centered computing—Animation—Virtual Characters—Virtual Reality;

1 Introduction

Nowadays, collaborative virtual environments (CVEs) allow people to communicate and collaborate with friends over a distance, form new relationships with strangers, and build visual embodiment through avatars. However, generating full-body motion for avatars [3], given only head and hands positions provided by consumer head mounted displays (HMDs), remains an ongoing problem. Most virtual environments only provide users with floating hands and heads, while a few can generate body motion through an inverse kinematics (IK) method, e.g. [9], which is not persuasive at all times. The lack of satisfying user motions forces us to think about whether this hinders people from collaborating in 3D telecollaboration systems and how we could achieve better user full-body motions.

The CoolMoves method [3], paves a way for us to utilize the motion capture (MoCap) dataset and real-time tracking data. The animations generated are more stylized and personalized than using static human animations or the IK method. However, the previous realization of CoolMoves is built upon the CMU MoCap dataset [1] which has a limited range of motions such as climbing and punching. Moreover, the previous evaluation only focuses on self-avatar embodiment, observed from a first-person perspective.

This project reports the preliminary work done to enhance participants' experience in CVE scenarios through full-body user motion accentuation. The purpose of this study is to construct a lightweight user motion generation method that has the potential to enhance interactions between users in CVEs by generating socially communicative gestures. In particular, we are interested in pointing toward targets so that an Observer can identify the correct target. The hypothesis is that by using our SocialMoves algorithm, a refined CoolMoves [3] algorithm, user communication will be more effective than an IK-based solution. To enable this situation we captured a new pointing dataset, PTMoCap on which to train the SocialMoves algorithm. In the pilot study, we haven't found a way to solve humans' lack of ability to interpret deictic gestures precisely in VR [8]. Therefore, we do not ask the Observer to call out the targets. In the future, we would conduct more comprehensive experiments.

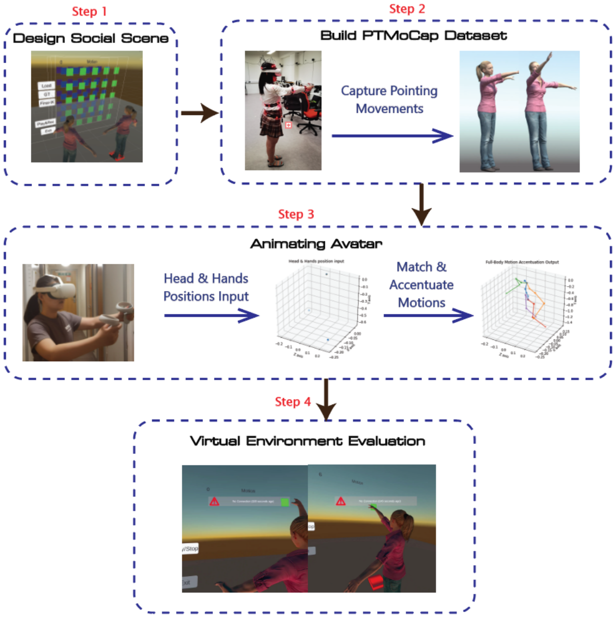
Step 1
Design Social Scene
Step 2
Build PTMoCap Dataset
Capture Pointing Movements
Step 3
Animating Avatar
Head & Hands Positions Input
Match & Accentuate Motions
Step 4
Virtual Environment Evaluation

Figure 1: Project Pipeline

1.1 Pipeline Design

Our project pipeline is shown in Figure 1 and contains four steps. Step 1: Design and implement a Social Pointing Scenario (SPS) which contains: 1) an Observer and a Pointer that supports various animating methods. 2) three walls, each with 25 controllable targets. Step 2: Design and capture movements in PTMoCap that can support right-arm pointing motions. Step 3: Implement the SocialMoves algorithm on PTMoCap. Step 4: Evaluate the method's applicability in SPS by comparing it with Final IK and Ground Truth.

1.2 Social Scene Design

We built an SPS that consists of two participants, each doing one simple activity: Observer A watches Pointer B pointing while being free to walk around. Pointer B points at different targets. We captured the motions of Pointer B to build our dataset. In 1994, Isaacs and Tang [5] conducted a systematic comparison of audio, video, and face-to-face as mediums for communication. They noted the great value of pointing at things in a shared environment. This previous study also gives us guidance on covering the set of possible pointing postures.

1.3 Pointing MoCap Dataset Generation

The capture system was a PhaseSpace [2] which uses active LED markers. In this project, we mainly focus on the action of pointing

*e-mail: ruoxiguo@126.com
†e-mail: l.izzouzi@ucl.ac.uk
‡e-mail: a.steed@ucl.ac.uk

Used Motion Classes in PTMoCap		
Motion Class	Posture	Number of Movements
1	Look Forward. Stand Still.	75
2	Rotate 45 degree left. Stand Still	75
3	Rotate 90 degree left. Stand Still	75
4	Walk around freely	75
5	Walk around freely	25

Table 1: PTMoCap Dataset Classes

with the right arm so only the upper-body joints were captured. The LED marker covered the head, chest, waist, right shoulder, right forearm (below elbow), and right hand. We recorded one subject doing 5 groups of pointing motions for a total of 870 seconds. This made up the self-captured PTMoCap dataset, shown in Table 1. The fifth motion class is used for testing the accentuation process.

1.4 Motion Accentuation

We implemented the SocialMoves algorithm generally following the ideas of CoolMoves [3] on Python 3.9. We first matched the feature map provided by the users' head and hands positions to the motions in PTMoCap. Then we obtained the smoothed weighted estimated full-body rotations and positions from the K-nearest matched motions. Challenge came when we tried to estimate the position of the root bone, and hip. Only knowing the head position, we undertook two different routes: 1. Use the average hip-head offset scalar estimated from the dataset as the offset between accentuated head and hip. 2. Use the offset calculated from the estimated hip and head positions. The latter has shown better animation results.

The accentuated animation is very accurate in movements. Compared with the ground truth, it successfully generated all 15 pointing movements with 2 unexpected movements of stretching the arm behind during the transition.

2 Evaluation in Social Virtual Environment

2.1 Experiment Design

During the evaluation process in SPS, the immersed participants observe an avatar's pointing motions in three different types of full-body animations: 1. Ground Truth. 2. Final IK solution. 3. Our Accentuated Motion method. We counterbalance the animation methods used. Our research goal is to validate our method. In our study, we use Rocketbox avatars to represent both participants. The Final IK employment on RocketBox Avatar in the Ubiq system is made possible through the method proposed by Izzouzi and Steed [6]. Each animation is 70 seconds in length and contains 15 movements of one avatar pointing at the appearing and disappearing targets (Figure 1, Step 4). Two female and two male volunteers took part in the pilot experiment. This study was approved by University College London Research Ethics Committee.

2.2 Evaluation Method

We measured users' social presence and Embodied intersubjectivity. The former was evaluated through partial Harms and Biocca's Social Presence (HSP) questionnaires [4] and the latter through a self-designed questionnaire. It is inspired by the discussion section on the distance between partners and the interactional circuit phases proposed by James [7]. The questionnaires, all in 1-7 rating scale, were conducted within 10 minutes after watching the full set of animations. The list of questions is attached in the supplementary material.

2.3 Evaluation Results

Social Presence Social presence in our questionnaire consists of co-presence, attentional allocation, and perceived message understanding. Our accentuation method (mean: 5.75 and 3.75) showed the same level of co-presence and perceived message understanding or even better than the ground truth (mean: 5.67 and 2.67), but not as good as the Final IK (mean: 6.33 and 4.17). For attentional allocation, the Final IK method has taken the most attention from our users (mean: 4.125), followed by Ground Truth (mean: 3.375) and Accentuation (mean: 3). In Perceived Message Understanding, participants are asked about how they understand the behaviors of their partner, the Pointer. Generally, they can tell that the character is trying to interact with the target board.

Embodied Intersubjectivity Similar to social presence, The Accentuation method (mean: 5.33) showed better results than the Ground Truth method (mean: 5.17), and slightly worse than the Final IK method (mean: 5.83). Participants who tried to click and point at the targets showed higher intersubjectivity scores.

3 Conclusion

This project was undertaken to enhance participants' experience in CVEs. We first captured a novel pointing dataset PTMoCap. Then, we implemented SocialMoves to generate full-body user motions from the user's hands and head positions through PTMoCap. We evaluated the accentuation algorithm's effect on social presence and embodied intersubjectivity in a social pointing scenario. Results show that the novel accentuated motion technique can generate accurate movements. It has the same level of performance as the ground truth, while Final IK scored a bit higher than the other two methods with extra lower body movements. This study shows the potential of data-driven full-body motion generation methods in social virtual reality. In the future, we will extend the accentuation method to integrate additional interactive social activities.

Acknowledgments

This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 860768.

References

[1] CMU MoCap, 2004. Retrieved December, 2022 from `http://mocap.cs.cmu.edu/`.

[2] Phasespace, 2022. Retrieved December, 2022 from `https://www.phasespace.com/`.

[3] K. Ahuja, E. Ofek, M. Gonzalez-Franco, C. Holz, and A. D. Wilson. Coolmoves: User motion accentuation in virtual reality. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 5(2), June 2021. doi: 10.1145/3463499

[4] C. Harms and F. Biocca. Internal consistency and reliability of the networked minds social presence measure. In *Seventh Annual International Workshop: Presence 2004*, 2004.

[5] E. Isaacs and J. Tang. What video can and cannot do for collaboration: A case study. *Multimedia Syst.*, 2:63–73, August 1994. doi: 10.1007/BF01274181

[6] L. Izzouzi and A. Steed. Integrating rocketbox avatars with the ubiq social vr platform. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 69–70, 2022. doi: 10.1109/VRW55335.2022.00025

[7] L. James. The collaborative construction of intersubjectivity mediated by technology: A social-biological model of online communication. *Symbiosis: SOJ Psychology*, 1:1–17, April 2014. doi: 10.15226/2374-6874/1/1/00114

[8] S. Mayer, J. Reinhardt, R. Schweigert, B. Jelke, V. Schwind, K. Wolf, and N. Henze. Improving humans' ability to interpret deictic gestures in virtual reality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (Honolulu, USA)*, pp. 1–14. Association for Computing Machinery, New York, NY, USA, 2020. doi: 10.1145/3313831.3376340

[9] D. Yang, D. Kim, and S.-H. Lee. Lobstr: Real-time lower-body pose prediction from sparse upper-body tracking signals. *Computer Graphics Forum*, 40(2):265–275, 2021. doi: 10.1111/cgf.142631