CoolMoves: User Motion Accentuation in Virtual Reality

KARAN AHUJA, Carnegie Mellon University, USA
EYAL OFEK, Microsoft Research, USA
MAR GONZALEZ-FRANCO, Microsoft Research, USA
CHRISTIAN HOLZ, ETH Zürich, Switzerland
ANDREW D. WILSON, Microsoft Research, USA

Current Virtual Reality (VR) systems are bereft of stylization and embellishment of the user's motion - concepts that have been well explored in animations for games and movies. We present *CoolMoves*, a system for expressive and accentuated full-body motion synthesis of a user's virtual avatar in real-time, from the limited input cues afforded by current consumer-grade VR systems, specifically headset and hand positions. We make use of existing motion capture databases as a template motion repository to draw from. We match similar spatio-temporal motions present in the database and then interpolate between them using a weighted distance metric. Joint prediction probability is then used to temporally smooth the synthesized motion, using human motion dynamics as a *prior*. This allows our system to work well even with very sparse motion databases (e.g., with only 3-5 motions per action). We validate our system with four experiments: a technical evaluation of our quantitative pose reconstruction and three additional user studies to evaluate the motion quality, embodiment and agency.

CCS Concepts: • **Human-centered computing** → **Human computer interaction (HCI)**; *Interaction Techniques*; User studies.

Additional Key Words and Phrases: Virtual Reality; Pose Tracking; Motion Embellishment; Motion Stylization.

ACM Reference Format:
Karan Ahuja, Eyal Ofek, Mar Gonzalez-Franco, Christian Holz, and Andrew D. Wilson. 2021. CoolMoves: User Motion Accentuation in Virtual Reality. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 2, Article 52 (June 2021), 23 pages. https://doi.org/10.1145/3463499

1 INTRODUCTION

Virtual Reality (VR) allows users to immerse themselves into imaginative worlds and enjoy appearances and behaviors that can be unlike anything in the real world. Users simply put on a Head Mounted Display (HMDs), pick up hand-held controllers, and enter a virtual environment where they can interact with the content around them. An important aspect of many VR experiences is the user's immersive perspective into the virtual world, which includes an avatar they embody and control through natural and continuous body movements [54].

The *first-person avatar* or *self-avatar* that represents the user typically follows the user's movements as performed. This *synchronized* mapping of the user's and the self-avatar's motions enhances embodiment (i.e., a sense of ownership of the virtual avatar [56, 63]), and the direct control of the avatar motions develops a sense of agency (i.e., a feeling of control over the avatar's actions and their consequences [50, 54]). Both are important characteristics of captivating VR experiences.

Authors' addresses: Karan Ahuja, Carnegie Mellon University, USA, kahuja@cs.cmu.edu; Eyal Ofek, Microsoft Research, USA, eyalofek@microsoft.com; Mar Gonzalez-Franco, Microsoft Research, USA, margon@microsoft.com; Christian Holz, ETH Zürich, Switzerland, christian.holz@inf.ethz.ch; Andrew D. Wilson, Microsoft Research, USA, awilson@microsoft.com.

2474-9567/2021/6-ART52
https://doi.org/10.1145/3463499

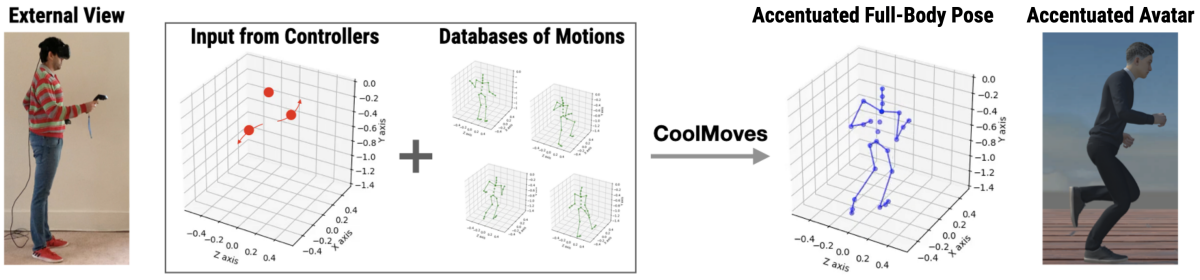

Fig. 1. From left to right: The motion of a user running (left) is only partially captured by a virtual reality system (middle), including the position and orientation of the head and two handheld controllers only. CoolMoves uses an existing motion capture database to match, accentuate, and blend the user's input into a full-body avatar motion. The resulting motion (right) thus resembles a professional athlete's motions, while the user maintains agency and embodiment. Note that in the external view, the user does not move their feet, but CoolMoves accentuates them based on the context.

The feel and control of VR experiences significantly differ from traditional interactive scenarios [72]. In console games, for example, avatars perform designed and highly stylized motion animations consistent with the game's genre. These high-quality animations are combinations of motion-captured sequences that skilled professionals performed in recording studios [64]. When playing a game, users can simply invoke such motions by the press of a button and perform elegant actions and easily transition between them. While these systems present the user with a sense of agency (control over avatar), there is no notion of embodiment (sense of body ownership), which is paramount to preserve given the inherent first-person nature of VR systems.

In this paper, we aim to bring these rich and sophisticated movements that make gaming experiences entertaining and desirable to VR scenarios, while maintaining the user's embodiment and agency. Unlike the real world, VR is not limited by physical laws, thus allowing the user to perform motions that are decoupled from the potentially unique and eloquent motions performed by their virtual avatar. We build on the approach of prior works that have disconnected the mapping between the user and the avatar in VR [39, 83] for reachability tasks. In this work, we expand on this disconnect and explore motion expressiveness and embellishment. As users of consumer VR lack the full-body motion capture and finesse to perform elaborate motion sequences, their avatar's movements can greatly vary from those performed by the non-player characters around them or other, more skillful players.

We propose *CoolMoves*, a Virtual Reality system that strikes a balance between the replay of pre-generated embellished animations employed in console games and a direct mapping of the user's motions to their virtual avatar. To the best of our knowledge, CoolMoves is the first system that implements user motion accentuation and stylization *in VR*, with a processing complexity that is effective for real-time use on commodity VR systems. CoolMoves dynamically synthesizes full-body movements and applies them to the user's avatar by fusing the user's actual motions with those drawn from a stylized repository containing examples recorded by professional skilled performers.

Subsequent interpolation between the user's input and candidate motions ensures that avatar motions strongly resemble the shape of the user's motion while maintaining the quality and spatial coherence of professional motion-capture animations. This interpolation also establishes a strong dependence of the synthesized movements to the input trajectories, such that CoolMoves can dynamically adjust to abruptly changing motions performed by the user. As a result, our system preserves full agency and embodiment for the user, but renders articulated and expressive movements for the avatar.

Animating and accentuating full-body motions accurately including the torso, hip, and feet is notably important in VR, as many games center around activities such as dancing, fighting, swimming, and so on. Despite the absence of actual tracking cues for most of the user's body, CoolMoves synthesizes consistent full-body animations purely from the hands and the head as shown in Figure 1. Our motion accentuation surpasses the full-body animations of current VR systems that derive avatar poses through inverse kinematics [2] (Figure 2). Note that the primary goal of CoolMoves is pose *accentuation*, not user pose estimation. This is evident in Figure 1 where the user does not move their feet but CoolMoves accentuates them. Thus, our system can not only accentuate the motions of the upper-body as seen in examples such as basketball and boxing in Figure 2, but can also generate the corresponding accentuated lower-body motions as seen in climbing and running. This full-body synthesis is inherent in our accentuation pipeline, but is not a precursor for it.

We evaluate CoolMoves with four experiments: a technical evaluation of our quantitative pose reconstruction on a public motion database, and three additional user studies to evaluate task congruence, embodiment and agency. We find that CoolMoves affords users fine control over a fully articulated, expressive avatar that can perform elegant motions while the user maintains embodiment and agency over the avatar throughout.

2 RELATED WORK

Our work sits at the interaction of graphics, perception, and virtual reality. Graphics research has long been understanding motion capture and motion stylization for animations and games, transferring motions between different skeletal structures [66] or even inanimate objects [31]. Human perception research has explored the phenomenon of agency [77] and body ownership [54, 60] as well as the effects that enhance or reduce them [62]. Particularly relevant are works that examined motion style transfer and those which have studied the discrepancies between the user's body and self-avatar in VR. We now briefly review these research areas.

2.1 Proprioception, Decoupling, and Body Ownership

Proprioception is the sense that enables humans to perceive where their body parts are in space. Our proprioception is normally in sync with our visual experience [40], but is dominated by the latter if they disagree. Through multi-sensory integration, researchers have shown that it is possible to produce systematic errors in proprioception towards another body [27]. This approach can be transferred to VR and the participant's proprioception can be decoupled from their body towards a virtual body [79]. The sense of ownership and agency over the avatar is generally known as *embodiment* [70].

An embodiment can be produced simply by being co-located with a virtual body [81], but it is enhanced by having full-body motions [55, 62]. In particular, [47] showed that motion-controlled VR avatars with full-body representation lead to an increased sense of presence, when compared to avatars that consist of only the head and hands. The human brain can accept some degree of mismatch between proprioception and the actual body location while performing motions and tasks [39]. This concept has been explored for spatio-temporal deformations of the virtual avatar [52], extending reachability [71, 84] and haptic retargeting among others [22, 49]. Broadly, redirection techniques exploit this proprioceptive decoupling [22, 32], however decoupling should not be too strong as it may break embodiment [28]. Prior research has found that the offset between the user's real and virtual hand can reach 14 cm before body ownership is reduced, and 21 cm before task performance degrades [84].

Along with body ownership, agency is a key component of embodiment [70]. Agency can be described as the feeling of control over one's own movements: I am the source of my movements [26, 54]. Studies by [35, 42] noted that the rapid adaptation of agency can occur even when avatars perform actions on different scales or speeds than the users, even if they are accompanied a proprioceptive drift [56]. However, if the avatar were to move autonomously, without intent of the user, the sense of agency would degrade. CoolMoves takes these

findings into account and generates motion of the virtual avatar that adapts to the temporal changes such as speeding, slowing down or change of direction of the user's input. Thus, even though CoolMoves introduces a proprioceptive drift due to motion accentuation, it preserves the agency and embodiment, as it follows the intention of the user.

2.2 Motion Capture for VR

Bringing an avatar guided by user input to life requires sensing the position and orientation of the user's body with high accuracy and low latency. The quality of the avatar animation depends on the number of body joints captured and the technique used to transfer motions to all the avatar's joints [79]. State-of-the-art systems for motion capture (MoCap), such as Optitrack [11] and Vicon [12] use multiple cameras to track wearable markers

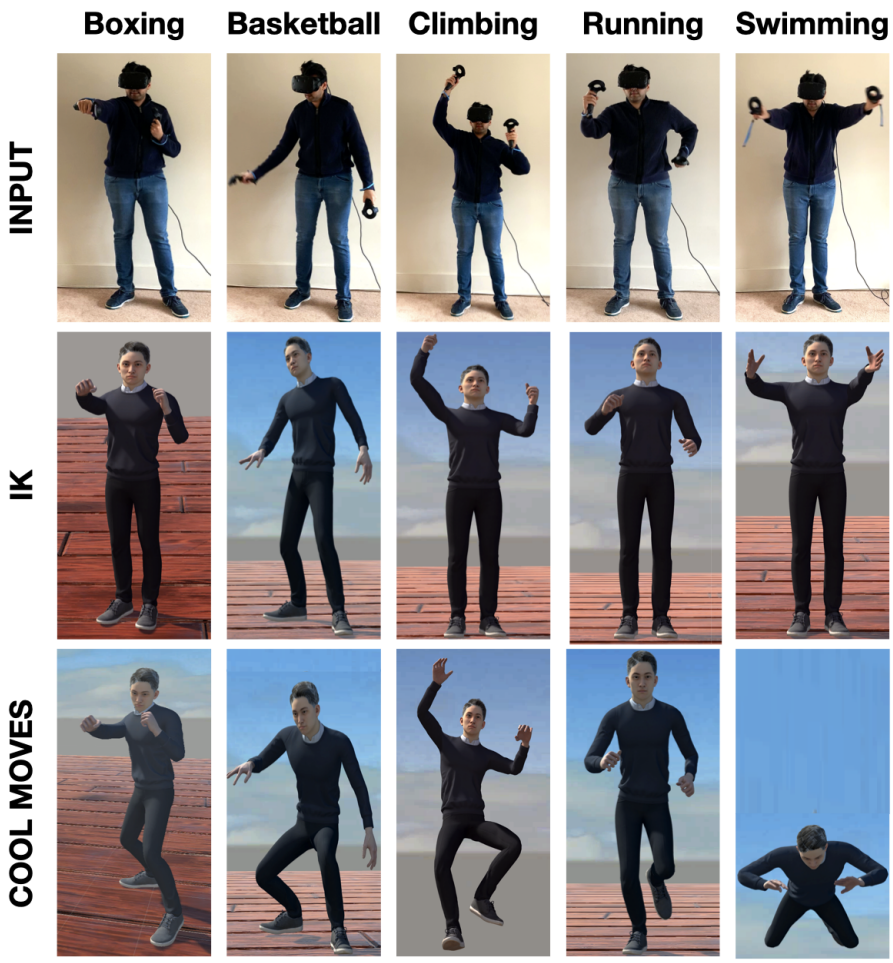

Fig. 2. Top row: External view of the user performing activities in VR, captured via the headset and two handheld controllers. Middle row: Output of IK baseline. Bottom row: Output of CoolMoves showcasing motion accentuation.

on the user's body [61]. Other technologies range from active wearable sensors such as inertial measuring units (IMUs) [46, 59] and cameras [76] to mechanical suits [7]. Using computer vision it is possible to track users motions without markers using depth sensors [5] or RGB cameras [17, 53] as long as the user is visible to the sensors.

However, current commodity VR systems such as HTC Vive [4], Oculus [10] or Windows MR [6] are limited to tracking the user's head and hands via the hand-held controllers and head mounted displays (HMDs). Such sparse sensing leaves VR applications to either render just the tracked joints or estimate the remainder of the body using Inverse Kinematics (IK) [2, 51, 69, 80] based approaches. Furthermore, approaches that focus on the physical appearance of the avatar, e.g. Social VR [57, 68] and co-located multi-player games [16] are bereft of any character motion style and embellishment. In contrast to these techniques, CoolMoves not only synthesizes the full-body motion from limited input in a data-driven manner but also accentuates it.

2.3 Stylized Motion Synthesis

The concept of motion "beautification" is well explored in domains such as sketching [23, 30, 73], dancing [34], motion capture [58], human-robot interaction [41] and motion-based gaming (e.g., Nintendo Wii games). A common approach for stylized motion synthesis employed by non VR games and character animation is to gradually transition between two motions given a certain direction, such as path or course control. Kovar et al. [58] and Rose et al. [74] explored the generation of different motion styles along arbitrary paths by smoothly blends motions between parameter control points, and [36] extends this to interactive character control. In contrast, CoolMoves has no foot contact or path information available to it. Furthermore, unlike transition based approaches where the first and next state are clearly demarked, VR presents a continuous user input.

Other style transfer techniques focus on parameterizing the spatio-temporal differences between the source and target style motions [20, 38, 78]. Data-driven methods for style transfer [85, 87] use existing MoCap databases for pre-defined motion types. Neural networks are used to model the "style" of motions and re-project it onto a destination subject [48], assuming full-body capture and may require offline motion cleaning and annotation. [13] makes use of GAN's to further extend these approaches to draw the style automatically from videos. In contrast to neural networks that require full-body data representations, CoolMoves generates a real-time full-body motion from limited motion sensing of just the head and hands. Furthermore, it is much less training data intensive and can interpolate motions from relatively sparse search spaces (at times as few as 6 motion instances for a style).

Currently, games such as FIFA [3] and NBA2K [8] have a big emphasis on the motion profiles of the characters along with their physical appearance. However, unlike console games that only deal with agency (control over avatar), VR systems need to preserve agency as well as embodiment (sense of ownership). Thus, the virtual avatar cannot be too far away the user's actual position (especially for the tracked joints) as this would break embodiment [39, 84]. CoolMoves was designed to balance between generating pleasing motions and maintaining full agency and embodiment. To the best of our knowledge, there is no prior work on motion styling in VR.

3 IMPLEMENTATION

CoolMoves decouples the avatar's motions from the user's input motions by accentuating them before rendering onto the skeleton of the self-avatar shown in VR. We now go over the various parts of our implementation pipeline.

3.1 Pose Accentuation

Transforming a source motion to an accentuated target motion presents several challenges. There is a trade-off between generating a motion that represents the accentuated (or stylized) motion of the target perfectly, while retaining the uniqueness of the user's input motion. This challenge is further exacerbated when dealing with

Table 1. Motion classes used from CMU MoCap Dataset for motion accentuation for CoolMoves

Motion	No. of Subjects	No. of Trials
Boxing	8	14
Basketball	4	52
Climbing Ladder	4	7
Running	7	55
Swimming	3	11

systems where the user generates the source motion while simultaneously viewing the accentuated motion. As these motions are generated in real-time, with no forecast of the user's intention and in an environment that requires low latency such as VR, agency, and embodiment are paramount. Therefore, our goal is to generate a motion that is similar to the user's performed motion while smoothly evolving from one type of motion to another in real-time (90 Hz for VR systems).

Keeping these challenges in mind, we propose a data-driven solution that leverages the expressive stylized motions collected in existing motion capture (MoCap) databases. We first start by generating candidate matches between the user input (from the headset and two handheld controllers) and the full-body poses in the MoCap dataset. We then interpolate between these candidate motions while optimizing to preserve the users original motion profile to maintain agency and embodiment.

3.1.1 MoCap Database. The Carnegie Mellon University (CMU) Motion Capture (MoCap) dataset [1] is one of the first and most extensive public motion capture datasets. It consists of various actions performed by over 140 subjects captured using 12 Vicon infrared MX-40 cameras, each of which is capable of recording full-body motion at 120 Hz. These actions range from locomotion to sports and interaction between people. It has been used extensively as a repository for template stylized motion by prior research [48, 85, 87]. For CoolMoves, we extracted five classes of activity from the CMU MoCap dataset as shown in Table 1. Each activity is performed by several subjects with each subject performing various trials of the same activity class. Each trial represents a single MoCap session. For example, within the boxing motion class, a single subject can have different MoCap trials, such as jabbing, side hook, etc.

3.1.2 Parametric Representation of Motions. To enable efficient matching of motion trajectories in real-time, CoolMoves creates a unified parametric representation of all the motions in the MoCap dataset. First, each MoCap trial is re-sampled to the tracking frequency of the VR system. As most consumer grade VR systems (including ours) register the headset and the hands at 90 Hz, we chose the same for our implementation. After re-sampling, we convert all the motions for each trial in the dataset to a head coordinate system by subtracting the position of the head from all the body joints. Note we do not normalize for head rotation in this step, as the rotation of the head (head-forward vector) can be independent of that of the hands (or the chest-forward vector). Thus, rotation normalizing would create ambiguity in the cases where the head is facing in a different direction than the chest.

The goal of our system is to generate full-body accentuated motions from the input of the users' headset and hands. Therefore, it is important to select upper body motions that lead to a high probability of inferring lower body motions. We achieve this by correlating the motions of the hands with the feet and only keep the ones that have a high correlation. To calculate the correlation, we employ the Pearson's correlation coefficient [75], treating the hands and the feet as a 3-dimensional time series (X, Y, Z position) across the entire motion sequence. For hand-dominant activities, such as basketball and climbing we found a high average correlation between the two (> 0.5). However, for activities such as swimming wherein the feet motion can be mutually independent of the hand, we found the mean correlation to be less than 0.2 (SD = 0.18) across all the trials in the activity. Therefore, it

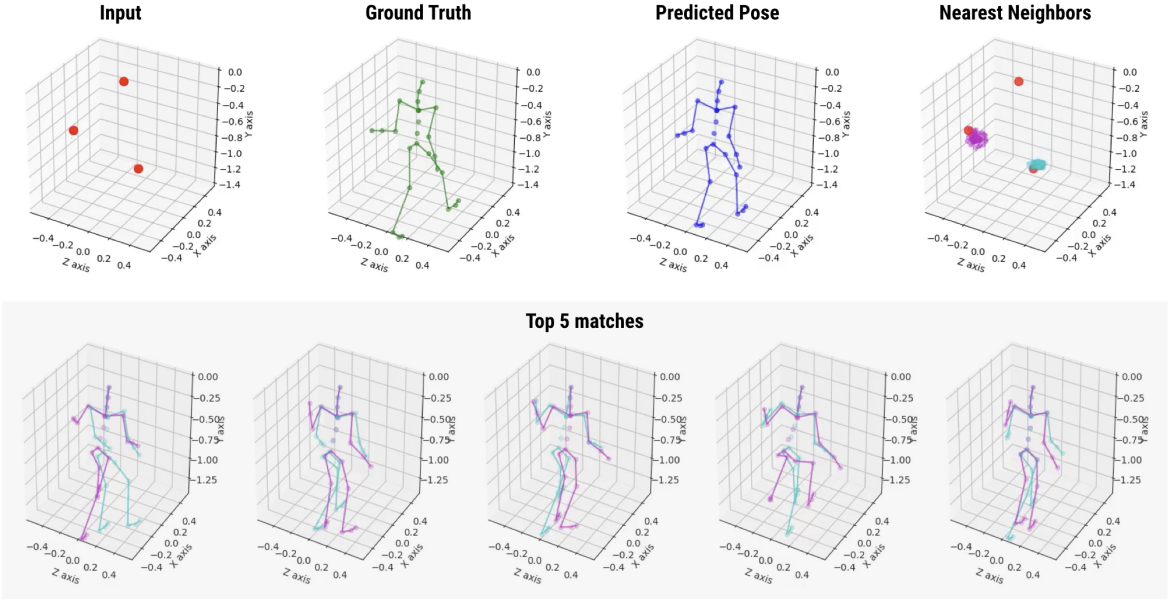

Fig. 3. Pose accentuation pipeline: The top row shows the generation of full-body accentuated poses from the input. The bottom row depicts the top 5 (out of total) matches to the input query for both the left (purple) and right hand (cyan).

becomes paramount to discard motions with a correlation lower than a threshold. This not only helps to decrease ambiguity during full-body pose reconstruction from the hands but also limits the search space and improves the computational efficiency. Since our algorithm can extrapolate from sparse motion space, we choose a relatively high threshold of 0.1. While this removes nearly 60% of the data from activities such as swimming, it ensures that motions present have a high chance of matching.

We now proceed to featurize the pose information for each activity, such that it captures the users motion profile. After applying the correlation threshold, we divide each MoCap trial using a rolling window of size 100 ms (9 samples of data) with an overlap of 50 ms between neighboring windows. Each window is a collection of full-body poses over time. However, since the input of our VR system is only the hands and head positions, we need to featurize each window to encode the same. This is achieved by first extracting the 3D positions of each hand from all the poses in the window and storing it into feature vectors corresponding to the left and right hand respectively. This helps encode the direction of motion from its start point. Note, we do not need to extract the head position as its already the origin. Subsequently, we concatenate the feature vector with the velocity and acceleration computed from its positional data. To simulate different speeds and intensities of motions, we further augment the dataset by generating multiple feature vectors and changing their velocity and acceleration profiles (up to 50% increase and decrease) followed by back-propagating these changes to the respective positions.

3.1.3 Matching Motions. Post featurization, for each MoCap motion, we have a set of rolling windows encapsulating full-body pose over time, and for each window we have feature vector corresponding to the left and right hand. For the matching stage, we first need to featurize the live input from the VR system. This is done similar to the featurization of the MoCap database, by first converting the input from the hands to a head-centered coordinate system and then converting the hand positions into a feature vector containing its acceleration and velocity profiles over a 100 ms rolling window.

Using the left and right-hand feature vectors, we estimated a mixture of Gaussians to model the probability of motions, similar to Xia et al. [85] and found it to be accurate but too slow for our needs. Hence we approximate the likelihood of a motion being applied for the motion synthesis by a function of matching distance. We, therefore, use a nearest neighbor approach to find a set of candidate matches as it is extremely light-weight and scalable to large datasets [21], and with almost same accuracy as Mixed of Gaussian methods. As the left and right hands have a high degree of independence and variability, we separately find the best candidate matches for each. To compute the matches we use a K-Nearest Neighbor [19] implementation based on a KDTree data structure with a neighborhood distance threshold, using the L2 norm as our distance metric. Figure 3 shows the matched left and right hand motions as purple and cyan dots, respectively, in the Nearest Neighbor plot.

3.1.4 Synthesizing Accentuated Motions. After the KNN matching step, the left and right hands provide us with k_l and k_r number of matched feature vectors, respectively, with a distance error for each match as well. Each feature vector corresponds to a 100 ms window of expressive full-body poses. We take the last pose from this window as the representative pose for synthesizing the accentuated full-body pose. This results in a total of k_l + k_r poses. The values of k_l and k_r can be tuned to balance the weighting between the user's input and the motions in the database.

Before interpolating between these poses, we first need to rotation align them to the input head up vector. As we do not have full-body pose upon input, we thus use the positional data of the hands in our feature vectors for alignment. We treat the 3D position of the hands as a temporal point cloud and find the relative rotation using a variant of Kovar's closed-form solution [58] (see equation). Since the rotation is along head-up (Y-axis), the optimization becomes:

$$\theta = \arctan \frac{\sum_i^n c_i(x_{q_i} z_{s_i} - x_{s_i} z_{q_i}) - \frac{1}{\sum_i^n}(\sum_i^n c_i x_{q_i} \sum_i^n c_i z_{s_i} - \sum_i^n c_i x_{s_i} \sum_i^n c_i z_{q_i})}{\sum_i^n c_i(x_{q_i} x_{s_i} + z_{s_i} z_{q_i}) - \frac{1}{\sum_i^n}(\sum_i^n c_i x_{q_i} \sum_i^n c_i x_{s_i} + \sum_i^n c_i z_{s_i} \sum_i^n c_i z_{q_i})}$$

Where c_i denotes the contribution of each of position in the feature vector, n represents the number of data points in 100 ms, q denotes the input feature vector, and s denotes the feature vector from the MoCap database. Here, x and z represent the x and z coordinates of the left and right hands. We assign the c_i from 0 to 1, increasing them in step-size $1/n$ from the start of the feature vector (0 ms) to the end (100 ms).

We now have k_l and k_r rotation-aligned poses corresponding to the matches from the left and right hand, respectively. To combine them, we first calculate the normalized weight (between 0 and 1) for each match, which is inversely proportional to the distance error. The proximity of a given hand to the respective body joints is directly proportional to the effect it has on it. In other words, the left and right arm's motion will be governed by their respective hand's motion, while the rest of the upper body and legs are common to both hands. Therefore, we only use the k_l left-hand matches to estimate the weights of the left arm joints (namely the hand, elbow, and shoulder) and use the k_r right-hand matches to estimate the weights of the corresponding right arm joints. For all the other joints, including the legs, hips, and torso we use a weighted average between the left and right arm weights. Using these weights for the poses, we interpolate between them to produce the output full-body pose. We also average out the weights of all the joints to compute a global match weight for the output pose. We find this approach to result in poses with output accentuated hands much closer to the original input hands.

We temporally smoothen the output pose matches between two consecutive windows. Motivated by prior research [15, 67, 82], we make use of an Exponential Weighted Moving Average (EWMA), taking the global match weight as the filter weight. Therefore, we have: $p_t = (w_t)p_t + (1 - w_t)p_{t-1}$ where p_t denotes the pose at time t and w_t is its corresponding global match weight. Other temporal smoothing methods have similar performances on the CMU MoCap dataset [67]. We then scale the bone lengths of the synthesized output skeleton to the match the user's height and add the global head translation of the VR headset. We further threshold the joint angles to constrain the degree-of-freedom for each joint based on human motion dynamics and to avoid self-occlusion. This results in accentuated full-body poses that are spatially and temporally coherent.

Matching the left and right hands independently also helps us better maintain agency and embodiment for the virtual avatar as the embellished joints are always relatively close to the tracked ones (in terms of position and cadence). Furthermore, there are motions that are predominantly focused on one hand (e.g. dribbling with one hand) or accessibility use cases wherein a single hand is only available for input. In such cases, matching the hands independently increases the fidelity of the accentuated motion.

We also note that while prior style transfer techniques for character animation have explored training data intensive deep learning based algorithms [13, 48], we opted for more lightweight approaches keeping the high speed and throughput constraints of VR (90 Hz) in mind. This is especially crucial when the GPU compute is either unavailable due to a standalone VR system or on a single GPU machine wherein the GPU is used for rendering the VR scene. Furthermore, these deep learning systems have relatively larger window sizes of 4 to 16 seconds of data, while the real-time nature of VR allows us to take a relatively smaller look-back window to minimize latency.

3.1.5 Processing System. Our runtime setup for CoolMoves consists of an AlienWare 15 R3 laptop with a Core i7 7700 HQ 2.8 GHz and NVIDIA GTX 1070 GPU. Our VR applications are developed in the Unity3D game engine, while the accentuation and MoCap processing backend run in Python. Our Gaussian mixture model [85] approach runs at 44 Hz, while our end-to-end nearest neighbor approach runs in real-time at 82 Hz. The Python backend for CoolMoves runs at 152 Hz, utilizing the CPU alone, suggesting that it could be used with VR systems that provide headset and controller positions at even higher frame rates.

4 EVALUATION

We conducted multiple users studies to systematically isolate and analyze different factors. In Section 4.1, we first benchmark CoolMoves ability to create coherent accentuated motions on a public motion capture dataset, which acts as our stylized template repository. This helps us evaluate whether the error bounds of CoolMoves lie within the range set by prior research to maintain body ownership. In the subsequent Sections 4.2 and 4.3 we evaluate the accentuated motion quality of CoolMoves on real-world data generated in a VR environment. Following this, in Section 4.4 we test the ability of CoolMoves to preserve agency and embodiment.

4.1 Pose Accentuation Evaluation

We run two quantitative evaluations to quantify the efficacy of our pose accentuation pipeline. The first is to measure the error of CoolMoves between the ground truth and generated accentuated pose to evaluate whether it remains within the bounds set by prior research to maintain embodiment and ownership. Second, we want to test the degree to which the virtual avatar follows the users motion profile. That is, if the user raises his hand up, the virtual avatar should also follow, albeit in a stylized manner. We calculate the correlation between the user's input and accentuated avatar generated by CoolMoves.

The CMU MoCap [1] database includes motions that are captured by multiple professional motion artists and therefore represent skillful and expressive motions. We make use of it [1] to evaluate our metrics. The database has motions across different activities and each activity has a variety of different motion trials within it, performed by multiple subjects. For our task, we choose a subset of five different activities: boxing, basketball, climbing, running and swimming (Table 1). We calculate the per-joint euclidean error and correlation on these activities and then further run an ablation study on it.

4.1.1 Evaluation Protocol. Unlike pose estimation which aims to minimize the reconstruction error between the full-body output pose and user's ground truth pose, the primary goal of CoolMoves is pose *accentuation*. This can be seen in Figure 2 where CoolMoves produces accentuated leg pose for activities such as boxing and running even though the input legs are static.

IK systems have been found to increase the sense of embodiment of consumer-grade VR systems [69, 80]. We thus make use of a state-of-the-art Inverse Kinematics (IK) as a baseline tool for pose synthesis from partial body tracking rather than full-body motion capture. For our IK baseline, we choose Unity's RootMotion Final IK [2] following [14, 69]. Specifically, we use the VR IK full-body IK solver module that supports tracking of the heads, hands, feet and body with corresponding bend goals and degree of freedom constraints, and inbuilt procedural locomotion. We use the IK module to generate the rest of the body based on the positions of the hands and the head. We test the performance of our algorithm and an IK baseline across five activities for two conditions: cross-subject and cross-trial.

In the cross-subject scenario, we use a leave-one-subject-out cross validation. Specifically, for each activity in Table 1, with number of subjects for that activity denoted by N_s, we use full-body data from $N_s - 1$ subjects as the motion repository and run CoolMoves and IK on the holdout subject. The holdout subject's full-body pose serves as the ground truth and its corresponding hands and head used as the input to CoolMoves to generate the accentuated full-body pose. Note, as all of these motions are drawn from the CMU MoCap dataset, they represent stylized and expressive motions. We run this across all holdout subject combinations and average the results. This result can be thought of as "out-of-the-box" accuracy, akin to running CoolMoves on a consumer-grade VR system such as HTC Vive which tracks the position of the headset and handheld controllers, but requires no calibration to the subject's motion profile.

In the cross-subject evaluation protocol, CoolMoves is never exposed to motion data from the subject it is accentuating. However, it is not uncommon for VR systems to collect some calibration data from the same user [18, 37]. The cross-trial evaluation protocol simulates this, wherein we performed a leave-one-trial-out cross validation. Specifically, for each activity with number of trials N_t, we use full-body data from $N_t - 1$ trial as the motion repository and run CoolMoves and IK on the holdout trial (all holdout trial combinations, results averaged). Similar to the previous evaluation protocol, CoolMoves estimates the accentuated full-body pose from this holdout trial's head and hand positions, which is then evaluated against the ground truth pose. As the stylized repository contains motion data from the subject we are accentuating, its akin to a per-user calibration model.

4.1.2 Results. Across the dataset, CoolMoves results in a mean error of 1.6 cm (SD = 0.22) for the left hand and 1.9 cm (SD = 0.27) for the right hand in the cross-subject scenario. In the cross-trial scenario, the error decreases to 1.4 cm (SD = 0.18) and 1.7 cm (SD = 0.21) for the left and right hand respectively. This shows that per-user calibration data helps improve system performance. The results of the total euclidean reconstruction error by body part configuration are shown in Figure 4 A. The upper-body encapsulates the joint of the head, neck, arms, and torso; and the lower-body encapsulates the joints of the pelvis, legs, and feet. CoolMoves has a mean full-body error of 11.4 cm (SD = 2.3) and 9.6 cm (SD = 2.1) for the cross-subject and cross-trial scenario respectively. These errors are well below the 14 cm threshold set by [84] to maintain body ownership. Our Gaussian Mixture Model-based approach, despite being slower, is slightly more accurate, with a mean full-body error of 10.6 cm (SD = 2.1).

Across different activities (Figure 4 B), we find that our system performs comparably across the board, apart for swimming. Upon further analysis, we found that swimming has an increased lower body error due to a limited correlation between the motions of the hands an the feet while the user swims, for example, you can swim with your feet or paddle them without much arm movement. We further run an ablation study, to see the contributions of the various parts of our algorithm. Using our nearest neighbor matching without any rotation alignment and probabilistic temporal smoothing, we get a cross-subject error of 16.5 cm across all activities. The error decreases to 13.2 cm upon the introduction of rotation alignment and, further to 11.4 cm when adding the temporal smoothing for consistency.

Our results are comparable to the IK baseline (see Figure 4 A) which has a mean full-body reconstruction error of 12.1 cm (SD = 3.4). The IK has no reconstruction error on the left and right hands as the solver starts optimizing

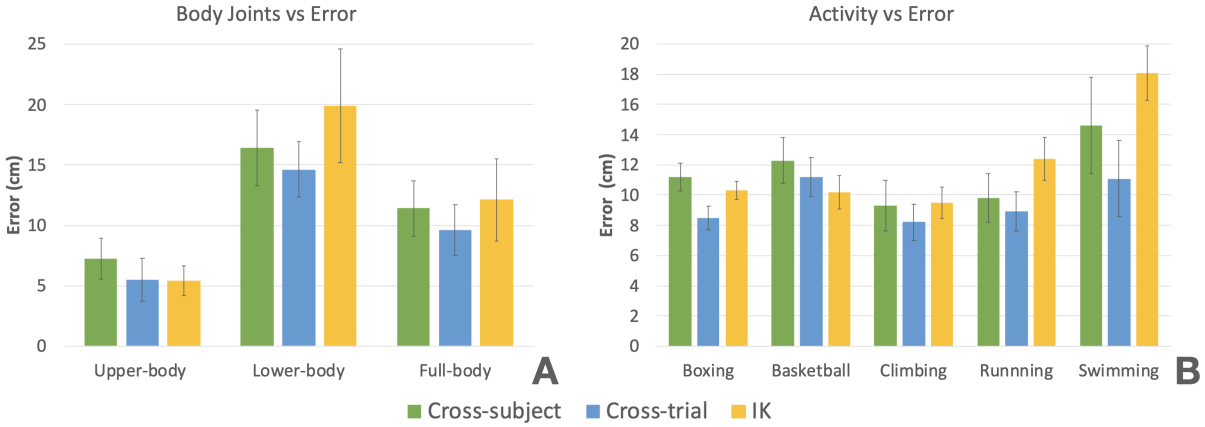

Fig. 4. A) The average joint error of CoolMoves for different subset of body joints configurations averaged across the dataset. B) The average full-body joint error of CoolMoves across 5 activities.

the inverse kinematic chain from these locations. This results in a smaller upper-body error. However, for the lower-body reconstruction, CoolMoves outperforms IK by 3.5 cm on average. This is expected, as IK lacks the activity-context and would default to a neutral position. This is evident in the accuracies of lower-body intensive activities such as running and swimming (see Figure 4 B). It is important to note that these comparisons are more of a reference guide, as the IK based systems are meant for reconstruction and not accentuation, which is the primary objective of CoolMoves.

To estimate the degree of correlation, we calculate the Pearson's correlation coefficient between the input hand positions and the output accentuated hands generated by CoolMoves. Specifically, we treat the X, Y and Z positions of the hands as independent time series and calculate the correlation coefficient for each of them, for the entire session (start to end of motion in the file). As the IK is set to the hand positions itself, it has a complete correlation of 1.0. CoolMoves achieves a high positive correlation of 0.92 and 0.94 for the left and right hands respectively (averaged across all three dimensions: X, Y and Z). Hence, we can conclude that the accentuated avatar's hands closely follows that of the input and should therefore help the user retain agency and embodiment over the accentuated avatar. We test this further in our subsequent qualitative user studies.

4.2 Motion Quality Evaluation: Crowd Study (no Agency)

Along with benchmarking the performance of CoolMoves on the CMU MoCap dataset, we evaluate its pose accentuation on real-world data collected from users in a VR environment. The goal is to judge the fidelity of stylization afforded by CoolMoves when the input comes from regular participants, who might not be as expressive as MoCap artists in their input motions. Such a system would not only be valuable for accentuating the motion of a user's avatar, but also lend in animating avatars for spectators or for third-person games in which the participant is not embodied in the avatar [45]. It would be especially useful in a social VR setting [68], collaborative work spaces [24, 25] or co-located multi-user VR experiences [16] wherein both the avatar motion and style need to be represented.

4.2.1 Procedure. For this study, we record motions performed by 5 experimenters using HTC Vive [4] VR headset and handheld controllers across our five activities: boxing, basketball, climbing, running and swimming. All five experimenters had experienced VR before with an average rating of 3.65 on a 7 point Likert scale. CoolMoves

(with CMU MoCap dataset as the stylized motion repository) and our IK baseline ran live in Unity and stored two rendered avatars. In line with prior work on pose stylization [85, 87], we use skeletal avatars rather than rigged character avatars. Skeletal avatars discount the effects of character mesh and the variable degrees of freedom afforded by different rigged models, thus focusing solely on the motion profiles for comparison.

Unlike in our pose accentuation study (Section 4.1), quantitative ground truth for the accentuated pose generated from the user's input does not exist. In line with prior work [25, 87], we conducted a crowd-sourced study on Amazon Mechanical Turk with 100 workers. Each crowd worker was paid $0.50 for the assignment and was asked the following 4 questions to assess the motion quality of CoolMoves:

(1) Which avatar better **captures** the style of the activity?
(2) Which avatar better **represents** the activity movement?
(3) Which avatar better represents the **upper** body motion?
(4) Which avatar better represents the **lower** body motion?

The above questions were inspired by prior studies on character stylization and style transfer [25, 87] wherein users answered questions related to style realism, representation to activity and comparison to prior techniques. We captured similar metrics but adapted them to the VR setting based on the avatar embodiment questionnaires by Peck et al. [70]. In the above questions, the definitions of upper and lower body were in line with our pose accentuation study (see Section 4.1.2). Each crowd worker was only shown one video (with two skeletal avatars: CoolMoves in blue and IK in white, with an average video length of 12 seconds) to mitigate learning artifacts, resulting in 20 crowd workers (× 4 questions = 80 responses) for each activity.

4.2.2 Results. Overall, a large percentage of the participants found the motions generated by CoolMoves represented the activity better (81%) and captured the style of the activity better (79%). The positive responses were not only predominant for the upper body joints (82%) but also for the lower body joints (75%) (Figure 5).

We further analyzed the results for these questions by fitting a generic linear model with mixed effects (glmer in the lme4 package, R), using a binomial distribution (of the logit family) that responded well to the need of analysis of repeated binary measures [29]. Then we ran an ANOVA Type III Wald χ^2 test on the model to explore the levels of significance. Question by question, results showed a main effect of condition (accentuated or non accentuated), $\chi^2(1) = 62, p < 0.0001$. The post-hoc test showed significant pairwise differences between conditions with significant higher scores for the accentuated avatars in these activities: Boxing (p=0.0001), Climbing (p=0.02), Basketball (p=0.003), Running (p=0.01). No significant differences were found for Swimming for any of the questions (p=0.06). The **captures** and **upper** questions also showed an interaction between condition and activity $\chi^2(4) = 12.1, p = 0.01$, beyond the main effect of condition.

CoolMoves outperformed the IK baseline for both upper and lower body accentuation (Figure 5). These results are in line with those from our quantitative pose accentuation study (Section 4.1). Thus while CoolMoves closely follows the input from the user's tracked joints (hands and head) to maintain embodiment, the generated accentuated full-body motion is visually distinct when compared to the state-of-the art IK, looking more natural and fitting the activity of the user.

4.3 Motion Quality Evaluation: Third person (with Agency)

Agency, or the sense of control, is generated by the timely reaction of the avatar motion to the actions of the user. While the location and velocity of the body joints can only be sensed by proprioception (awareness of the position and movement of the body) at limited accuracy, it is sensitive to temporal changes, such as motion direction change, or acceleration/deceleration of motions.

We ran a real-time experiment to focus on agency and perceived quality during live control of CoolMoves. Note, in this setup, there is no embodiment of the avatar despite it is being controlled by the participant. We focus solely on the agency of the motions and the kinematics. This study helps dissociate how much agency

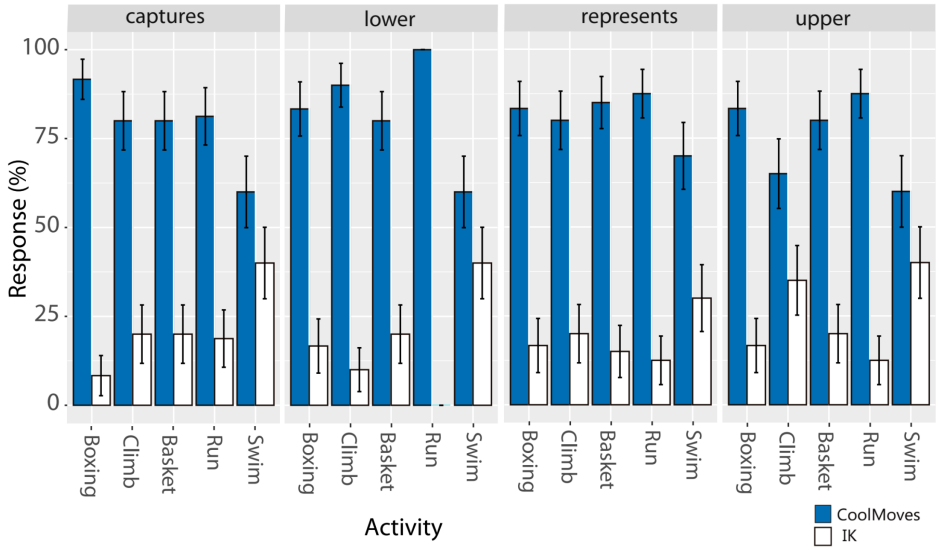

Fig. 5. Crowd-sourced study results showing the % of responses for CoolMoves and IK avatar respectively for different activities and questions asked in the experiment.

vs proprioception is distorted by CoolMoves. Besides checking the agency this study is also a good proxy to understand how good a third person rendering will be with CoolMoves. On certain occasions, the user may see the avatar from other points of view, such as in a mirror or a shadow, and it is important that this third person point of view will be consistent with the user's motion too.

4.3.1 Participants. 12 participants (8 male, 4 female) volunteered for our experiment, their ages ranged from 25 to 35, with a mean age of 28.6 years (SD = 3.2). The average study session lasted for 22 minutes. All participants were right-handed, with a mean height of 1.75 meters and had experienced VR before with an average rating of 3.25 on a Likert Scale of 1 to 7.

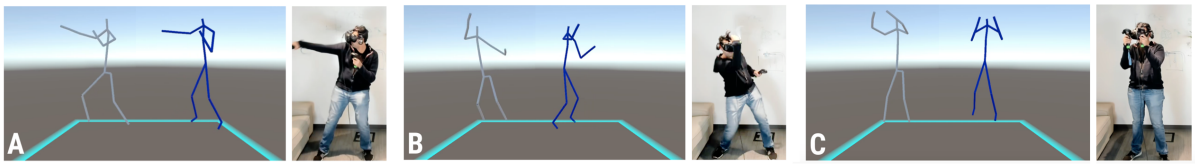

Fig. 6. Comparison of the user's motion between IK (white) and CoolMoves (blue). The pose produced by CoolMoves is better accentuated towards boxing maneuvers. While punching and performing an upper cut in A and B respectively, CoolMoves better captures the style of boxing: the legs generated via CoolMoves move to support the body weight while those generated by IK stay at the same place. Difference in the elbows and forearms is also visible, especially for the hand that is punching. This difference is stark while blocking the face in C, where CoolMoves uses the full forearm to guard the face.

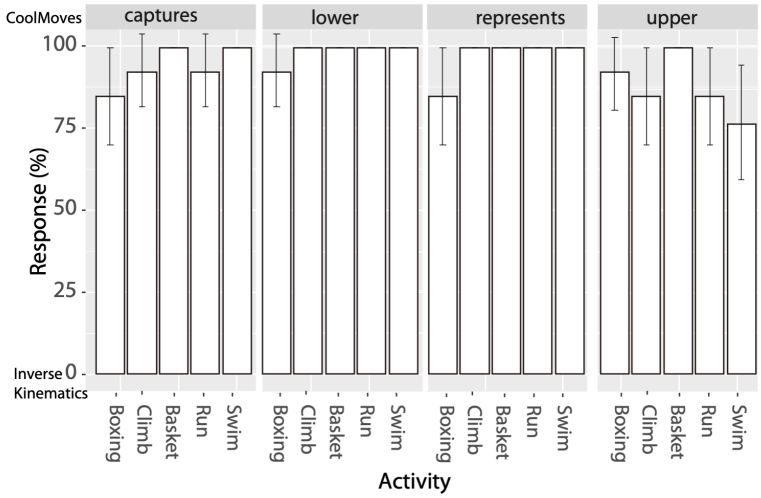

Fig. 7. Results for agency evaluation. Active observer: third-person motion accentuation quality measure study. Preference of CoolMoves vs IK (vertical axis) across different questions.

4.3.2 Procedure. In this study, we asked participants to perform motions corresponding to five different activities: boxing, basketball, climbing, running, and swimming (activity sequence ordering was randomized for each participant). While performing the motions they saw two skeletal avatars in front of them, which they saw in third-person inside the HMD. One of the avatars was rendered using real-time CoolMoves results and the other one was using Inverse Kinematics (IK) baseline based on Unity's Final IK [2]. Each participant was given a minute each to get a feeling of each of the five scenarios. Then they were told and guided to perform multiple motions depending of the activity: boxing, basketball, climbing, running, and swimming. Each activity session lasted for about 3 minutes. An example illustration for the boxing activity can be seen in Figure 6. After each activity, participants responded to the same 4 questions as those in the crowd study (to showcase the difference between a passive and an active observer). These questions are relevant to agency as well because they talk about representing and capturing the motion quality of the avatar movements generated as a direct result of the user performing them. Consequently, this also helps us evaluate the effect of agency on the perceived motion quality of CoolMoves.

4.3.3 Results. The results (Figure 7) show a clear preference of participants towards the CoolMoves avatar. Across all questions, 94.4% (SD = 0.17) of the participants preferred the motion quality of the CoolMoves avatar across all activities. CoolMoves outperformed the IK baseline for both upper (87.7%) and lower body (98.4%) accentuation. 93.8% and 96.9% felt that CoolMoves represented and captured the style of the activity better. This preference is even more pronounced than seen in the crowd-sourced study. We hypothesize that this is due to the additional agency that users experienced in this experiment vs. the observer crowds. It can also be attributed to the increased display capability - participants could see the motion in stereo and with head motion parallax, which may have made the difference even more pronounced.

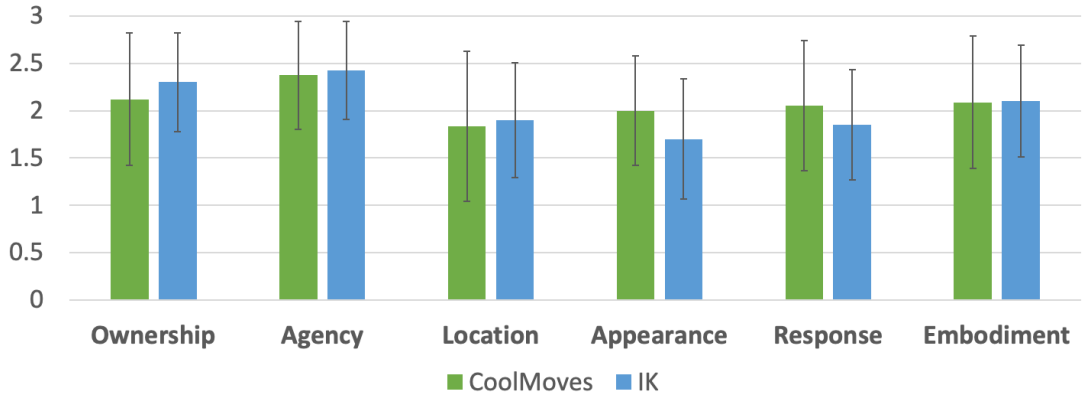

Fig. 8. Ownership, agency, location, appearance, response and total embodiment scores of CoolMoves vs IK baseline.

4.4 Embodiment Evaluation

Embodiment can be defined as the sense of ownership of the virtual avatar [56, 63]). We evaluated CoolMoves on boxing as the embodiment scenario (Figure 9). In this study, participants go beyond controlling an avatar in third person, but they are shown this avatar from a first-person view, such as if the avatar is representing the participant's body in VR.

4.4.1 Participants. We recruited 10 participants (7 male, 3 female) for this study (new participants, independent of the prior studies). The average height was 1.52 meters and the average age was 23.4 years (SD = 2.1). All participants had experienced VR before, with an average rating of 3.4 on a Likert Scale of 1 to 7. The average study session lasted for 25 minutes.

4.4.2 Procedure. In this experiment, participants experienced 8 minutes of boxing in front of a mirror and against a boxing bag while embodied in the avatar that was driven by CoolMoves. After a 2 minute break, 8 more minutes of boxing in the same scene followed, but with a avatar driven by our IK baseline. The order between CoolMoves and IK was alternated for every other participant. The participants also had no knowledge of whether CoolMoves or IK was being used to drive their avatar. At the end of the experience, they responded to an embodiment questionnaire [43]. In particular, the embodiment questions were on a scale of -3 to 3 with their corresponding score metrics as follows:

(1) **Body Ownership:** (A) I felt as if the virtual body was my own. (B) I felt as if the virtual body I saw was someone else. (C) It seemed as if I might have more than one body. (D) I felt as if the virtual body I saw when looking in the mirror was my own body. (E) I felt as if the virtual body I saw when looking at myself in the mirror was another person.
$Ownership = (1A - 1B) - 1C + (1D - 1E)$
(2) **Agency:** (A) It felt like I could control the virtual body as if it was my own body. (B) The movements of the virtual body were caused by my movements. (C) I felt as if the movements of the virtual body were influencing my own movements. (D) I felt as if the virtual body was moving by itself.
$Agency = 2A + 2B + 2C - 2D$
(3) **Location of the body:** (A) I felt as if my body was located where I saw the virtual body. (B) I felt out of my body. (C) I felt as if my (real) body were drifting toward the virtual body or as if the virtual body were drifting toward my (real) body.

$Location = 3A - 3B + 3C$

(4) **External Appearance:** (A) It felt as if my (real) body were turning into an 'avatar' body. (B) At some point, it felt as if my real body was starting to take on the posture of the virtual body that I saw.
$Appearance = 4A + 4B$

(5) **Response to external stimuli:** (A) I felt that my real body could be affected by the virtual body. (B) I felt as if my real body had changed.
$Response = 5A + 5B$

We removed the questions related to tactile sensation, apparel and bodily threat from the general questionnaire as they were not applicable to our experiment. Following [43], the gross embodiment score can be represented as $TotalEmbodiment = ((Ownership/5) * 2 + (Agency/4) * 2 + (Location/3) * 2 + (Appearance/2) + (Response/2))/8$

4.4.3 Results. The results across the different embodiment metrics averaged across all participants are shown in Figure 8. The results show that there were no significant differences in the reported total embodiment between CoolMoves (mean score: 2.08) and the IK (mean score: 2.1) baseline. This is a promising result as the accentuated motion will tend to lead to a decrease in embodiment due the mapping of the joints to a different position than the original ones. CoolMoves ability to retain the user's intent and motion profile can be seen in the marginal difference in the agency scores between CoolMoves (2.37) and the IK baseline (2.42) - a difference of only 0.05.

Our results indicate that it is possible to have a data-driven approach to generate real-time accentuated avatar motion even in first-person embodiment scenarios. While they are promising, it is important to note that the lack of significant difference between the embodiment scores of CoolMoves and IK might be due to the limited size of our participant pool thereby diminishing the statistical power of the samples.

5 APPLICATIONS

To demonstrate the utility of CoolMoves, we describe a variety of illustrative use cases that could be incorporated into VR applications to enhance embodiment and immersion. Please also see the supplementary video figure.

Fig. 9. First-person view of the person boxing by person in VR for the embodiment evaluation.

Fig. 10. CoolMoves uses limited input of the user's hands and head to recover plausible legs motions as showcased in this scene of a user climbing a ladder.

5.1 Accentuated Full-Body Avatars

Sensing in today's consumer VR is limited to the hand and head tracking, thereby immensely restricting the range of motions of the user. Lower body, if at all generated, is often animated by controllers or naively follows the global position of the user.

Our data-driven approach in CoolMoves can leverage the intrinsic correlations between body parts to enable a richer form of interaction and immersion. The fact that parts of the body act in a synchronous way can be used to provide VR characters with full-body avatar renderings that imbue not only the physical appearance but also the style of the avatar. For example, the effects of the synchronized feet motion with the hands can be seen while the avatar is climbing (Figure 10), or serve to reconstruct leg motions during walking or running (Figure 1). Note, that the lack of movement of the legs is such cases is not due to the lack of sensing, but rather the users do not move their legs (see Figure 1 and 10). CoolMoves generates this motion based on the context and movement of the hands. Furthermore, views of the self-avatar in reflections (such as in mirrors) and shadows that follow the user's motion increase the game embodiment [44].

5.2 Affect and Emotion on Avatar Motions

In a social VR setting or in multi-player games, CoolMoves can be used to imbue users emotion on their avatars. This can be done by selecting different motion styles for their avatars in real-time (Figure 11). Affect read through other sensors can then be easily transferred to the avatar. Furthermore, in applications such as games, CoolMoves could be used to convey the state of the character rather than the user, such as being drunk, bitten, injured or tired.

5.3 Limb Exaggeration and Scaling

Moving in VR gives a great sense of agency. However, many times after some period of use, there is a need to sit down, sometimes due to fatigue, or room size constraints. Recent work has looked at scaling motions to reduce fatigue [84]. CoolMoves can be combined with a range of transformations such as scaling to exaggerate motions of player, thus decreasing the overall motion footprint, while preserving immersion and embodiment (Figure 13). Apart from scaling, CoolMoves can help in the remapping of limbs. For example, the user can make use of their hands to run while sitting on a chair (see Figure 12). This capability can enable interesting accessibility applications for VR in the near future.

6 LIMITATIONS AND FUTURE WORK

CoolMoves makes use of the input from only the headset and hand-held controllers. It therefore cannot directly sense the legs. This leads to jitters in lower body reconstruction as well as foot sliding artifacts. Furthermore, its accuracy will degrade in motions that are primarily foot-driven such as soccer. In the future, these challenges can be overcome by sensing the feet using additional hardware and adding their features to the nearest neighbor search.

In the current implementation of CoolMoves, the stylized motion dataset consists of a small set of professionally recorded motions of the same context. Whenever a user moves in a way that is very different from the motions in that category of the data set, our implementation will converge on a one-to-one mapping to the user's input motions. We plan to generalize our matching approach beyond individual categories to predict full-body locations independent of activity context in the future. Another interesting exploration along these lines is the interpolation of activity from a different activity class, for example, playing basketball while loading the stylized motion data from the running activity. Currently, in such cases, CoolMoves will superimpose the estimated motions, e.g. dribbling while running rather than retaining the motion profile of the original activity being performed. While this might be beneficial in some use cases, the exploration of CoolMoves as a style transfer for human motion between independent actions remains as future work.

Different users may be more sensitive to variations in the avatar motion from their own body than others. By setting the weights, it is possible to restrict the motion synthesized by CoolMoves to be closer to the original motion while reducing the applied accentuation. Also, we do not currently apply anatomical limits on the motions applied to the avatar beyond the length of bones. In the future, CoolMoves can employ better motion retargeting techniques that can take different bone hierarchies into account as well. It can also be combined with foot IK or other physics based engines to improve the sense of agency and embodiment, for example in real-walking scenarios in VR [33, 86].

CoolMoves is currently designed keeping the embodiment and body ownership requirements of VR systems in mind. However, the high agency afforded by it, coupled with its low latency and real-time nature can also make

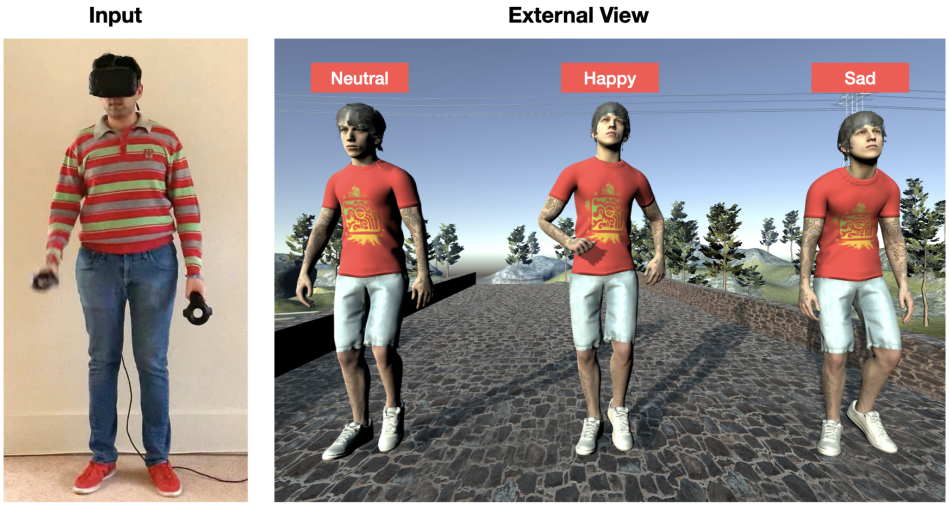

Fig. 11. Example scene showcasing how CoolMoves can add different affect to the user walking motion.

Fig. 12. Example scene showcasing limb remapping with the user running with their hands as input.

is useful as a general pose accentuation framework. In future, CoolMoves can be extended and evaluated for different input modalities such as IMU (e.g. Nintendo Switch [9]) and camera-based gaming (e.g. Kinect games [65]) systems across a wide range of motions including sports, dance, and role-playing games.

7 CONCLUSION

We presented CoolMoves, a real-time VR system that synthesizes expressive and fully articulated full-body motions for a VR avatar. CoolMoves generates such full-body poses using the positions of the user's headset and the controller positions alone. While CoolMoves' synthesized motions originate from a convolution of motion-captured demonstrations by professional performers, CoolMoves dynamically fits them to the user's actual body pose and motion. Therefore, our method affords the user full agency over the avatar while showing elegant and athletic body motions to the user.

Fig. 13. Motion Exaggeration with CoolMoves. Left: External view of the user boxing. Middle: Avatar rendered via CoolMoves with no exaggeration. Right: CoolMoves avatar with motion up-scaling on the input hands thus leading to an exaggerated avatar output.

Comparably expressive body motions have previously only been achieved by console games, which, however, only play back pre-recorded motion capture animations upon pressing controller buttons. CoolMoves, in contrast, processes 3D input and transforms motion trajectories on-the-fly and while such input motions are still happening, which establishes the desirable perception of embodying the VR avatar. We achieve this by continuously matching the user's motion trajectories against a large motion capture database of athletic motions, synthesizing a novel motion from the best matches, and applying it to the avatar's current trajectory.

In our user study, the accuracy of the motions synthesized by CoolMoves compared to those in the CMU Motion Capture database, which contains motions performed by professional actors. Our study also showed a preference in participants' ratings during first-person use and spectatorship for CoolMoves' synthesized avatar motions. Finally, we see CoolMoves as an almost necessary complement for current VR systems. Most consumer-grade VR systems can only track the users head and hands. As a result, animated characters appear with insufficient fidelity, especially about the motion of lower body parts. CoolMoves full-body motion synthesis restores much higher-fidelity kinetics to these body parts by leveraging a full-body motion database.

REFERENCES

[1] 2004. CMU MoCap. http://mocap.cs.cmu.edu/.
[2] 2018. RootMotion Final IK. https://assetstore.unity.com/packages/tools/animation/final-ik-14290.
[3] 2020. FIFA2020. https://www.ea.com/en-au/games/fifa.
[4] 2020. HTC Vive. https://www.vive.com/.
[5] 2020. Microsoft Kinect. https://azure.microsoft.com/en-us/services/kinect-dk/.
[6] 2020. Microsoft Mixed Reality. https://www.microsoft.com/en-us/windows/windows-mixed-reality/.
[7] 2020. Motion Capture - Meta Motion sells Motion Capture Hardware and Software. https://metamotion.com/
[8] 2020. NBA2K. https://nba.2k.com/2k20/en-US/.
[9] 2020. NintendoSwitch. https://www.nintendo.com/switch/.
[10] 2020. Oculus. https://www.oculus.com/.
[11] 2020. Optitrack. https://www.optitrack.com/.
[12] 2020. Vicon. https://www.vicon.com/.
[13] Kfir Aberman, Yijia Weng, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. 2020. Unpaired Motion Style Transfer from Video to Animation. *arXiv preprint arXiv:2005.05751* (2020).
[14] Parastoo Abtahi, Mar Gonzalez-Franco, Eyal Ofek, and Anthony Steed. 2019. I'm a Giant: Walking in Large Virtual Environments at High Speed Gains. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 522.
[15] Michael Adjeisah, Yi Yang, and Lian Li. 2015. Joint Filtering: Enhancing gesture and mouse movement in Microsoft Kinect application. In *2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*. IEEE, 2528–2532.
[16] Karan Ahuja, Mayank Goel, and Chris Harrison. 2020. BodySLAM: Opportunistic User Digitization in Multi-User AR/VR Experiences. In *Symposium on Spatial User Interaction*. 1–8.
[17] Karan Ahuja, Chris Harrison, Mayank Goel, and Robert Xiao. 2019. MeCap: Whole-Body Digitization for Low-Cost VR/AR Headsets. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. 453–462.
[18] Karan Ahuja, Rahul Islam, Varun Parashar, Kuntal Dey, Chris Harrison, and Mayank Goel. 2018. Eyespyvr: Interactive eye sensing using off-the-shelf, smartphone-based vr headsets. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (2018), 1–10.
[19] Naomi S Altman. 1992. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician* 46, 3 (1992), 175–185.
[20] Deepali Aneja, Alex Colburn, Gary Faigin, Linda Shapiro, and Barbara Mones. 2016. Modeling stylized character expressions via deep learning. In *Asian conference on computer vision*. Springer, 136–153.
[21] Ahmed Shamsul Arefin, Carlos Riveros, Regina Berretta, and Pablo Moscato. 2012. Gpu-fs-knn: A software tool for fast and scalable knn computation using gpus. *PloS one* 7, 8 (2012).
[22] Mahdi Azmandian, Mark Hancock, Hrvoje Benko, Eyal Ofek, and Andrew D Wilson. 2016. Haptic retargeting: Dynamic repurposing of passive haptics for enhanced virtual reality experiences. In *Proceedings of the 2016 chi conference on human factors in computing systems*. ACM, 1968–1979.
[23] Mayra D Barrera Machuca, Paul Asente, Jingwan Lu, Byungmoon Kim, and Wolfgang Stuerzlinger. 2017. Multiplanes: Assisted freehand VR drawing. In *Adjunct Publication of the 30th Annual ACM Symposium on User Interface Software and Technology*. 1–3.

[24] Hrvoje Benko, Edward W Ishak, and Steven Feiner. 2004. Collaborative mixed reality visualization of an archaeological excavation. In *Third IEEE and ACM International Symposium on Mixed and Augmented Reality*. IEEE, 132–140.
[25] Uttaran Bhattacharya, Nicholas Rewkowski, Pooja Guhan, Niall L Williams, Trisha Mittal, Aniket Bera, and Dinesh Manocha. 2020. Generating Emotive Gaits for Virtual Agents Using Affect-Based Autoregression. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 24–35.
[26] Olaf Blanke and Thomas Metzinger. 2009. Full-body illusions and minimal phenomenal selfhood. *Trends in cognitive sciences* 13, 1 (2009), 7–13.
[27] Matthew Botvinick and Jonathan Cohen. 1998. Rubber hands 'feel'touch that eyes see. *Nature* 391, 6669 (1998), 756.
[28] Sidney Bovet, Henrique Galvan Debarba, Bruno Herbelin, Eray Molla, and Ronan Boulic. 2018. The critical role of self-contact for embodiment in virtual reality. *IEEE Transactions on Visualization and Computer Graphics* 24, 4 (2018), 1428–1436.
[29] Jerry Brunner. 2019. Repeated measurement analysis of binary responses. http://www.utstat.toronto.edu/~brunner/workshops/mixed/
[30] Sonya Cates and Randall Davis. 2004. New approach to early sketch processing. *Making Pen-Based Interaction Intelligent and Natural* (2004), 29–34.
[31] Izadi S. Chen, A. and A. Fitzgibbon. 2012. KinÊtre: animating the world with the human body.. In *Proceedings of the ACM symposium on User interface software and technology (UIST 2012)*.
[32] Lung-Pan Cheng, Eyal Ofek, Christian Holz, Hrvoje Benko, and Andrew D Wilson. 2017. Sparse haptic proxy: Touch feedback in virtual environments using a general passive prop. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 3718–3728.
[33] Lung-Pan Cheng, Eyal Ofek, Christian Holz, and Andrew D Wilson. 2019. VRoamer: generating on-the-fly VR experiences while walking inside large, unknown real-world building environments. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 359–366.
[34] Worawat Choensawat, Sachie Takahashi, Minako Nakamura, Woong Choi, and Kozaburo Hachimura. 2010. Description and reproduction of stylized traditional dance body motion by using labanotation. *Transactions of the Virtual Reality Society of Japan* 15, 3 (2010), 379–388.
[35] Brian J Cohn, Antonella Maselli, Eyal Ofek, and Mar Gonzalez Franco. 2020. SnapMove: Movement Projection Mapping in Virtual Reality. In *AIVR 2020*. IEEE.
[36] Marco da Silva, Yeuhi Abe, and Jovan Popović. 2008. Interactive simulation of stylized human locomotion. In *ACM SIGGRAPH 2008 papers*. 1–10.
[37] Brian Day, Elham Ebrahimi, Leah S Hartman, Christopher C Pagano, Andrew C Robb, and Sabarish V Babu. 2019. Examining the effects of altered avatars on perception-action in virtual reality. *Journal of Experimental Psychology: Applied* 25, 1 (2019), 1.
[38] Ruta Desai, Fraser Anderson, Justin Matejka, Stelian Coros, James McCann, George Fitzmaurice, and Tovi Grossman. 2019. Geppetto: Enabling Semantic Design of Expressive Robot Behaviors. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 369.
[39] Tiare Feuchtner and Jörg Müller. 2018. Ownershift: Facilitating Overhead Interaction in Virtual Reality with an Ownership-Preserving Hand Space Shift. In *The 31st Annual ACM Symposium on User Interface Software and Technology*. ACM, 31–43.
[40] Shaun Gallagher. 2000. Philosophical conceptions of the self: implications for cognitive science. *Trends in cognitive sciences* 4, 1 (2000), 14–21.
[41] Michael J Gielniak, C Karen Liu, and Andrea L Thomaz. 2010. Stylized motion generalization through adaptation of velocity profiles. In *19th International Symposium in Robot and Human Interactive Communication*. IEEE, 304–309.
[42] Mar Gonzalez-Franco, Brian Cohn, Eyal Ofek, Dalila Burin, and Antonella Maselli. 2020. The self-avatar follower effect in virtual reality. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 18–25.
[43] Mar Gonzalez-Franco and Tabitha C. Peck. 2018. Avatar Embodiment. Towards a Standardized Questionnaire. *Frontiers in Robotics and AI* 5 (June 2018), 74. https://doi.org/10.3389/frobt.2018.00074
[44] Mar Gonzalez-Franco, Daniel Perez-Marcos, Bernhard Spanlang, and Mel Slater. 2010. The contribution of real-time mirror reflections of motor actions on virtual body ownership in an immersive virtual environment. In *2010 IEEE virtual reality conference (VR)*. IEEE, 111–114.
[45] Geoffrey Gorisse, Olivier Christmann, Etienne Armand Amato, and Simon Richir. 2017. First-and Third-Person Perspectives in immersive Virtual environments: Presence and Performance analysis of embodied Users. *Frontiers in Robotics and AI* 4 (2017), 33.
[46] Sehoon Ha, Yunfei Bai, and C Karen Liu. 2011. Human motion reconstruction from force sensors. In *Proceedings of the 2011 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. 129–138.
[47] Paul Heidicker, Eike Langbehn, and Frank Steinicke. 2017. Influence of avatar appearance on presence in social VR. In *2017 IEEE Symposium on 3D User Interfaces (3DUI)*. IEEE, 233–234.
[48] Daniel Holden, Ikhsanul Habibie, Ikuo Kusajima, and Taku Komura. 2017. Fast neural style transfer for motion data. *IEEE computer graphics and applications* 37, 4 (2017), 42–49.
[49] David Antonio Gomez Jauregui, Ferran Argelaguet, Anne-Helene Olivier, Maud Marchal, Franck Multon, and Anatole Lecuyer. 2014. Toward" pseudo-haptic avatars": Modifying the visual animation of self-avatar can simulate the perception of weight lifting. *IEEE*

transactions on visualization and computer graphics 20, 4 (2014), 654–661.
[50] Marc Jeannerod. 2002. The mechanism of self-recognition in humans. , 15 pages.
[51] Fan Jiang, Xubo Yang, and Lele Feng. 2016. Real-time full-body motion reconstruction and recognition for off-the-shelf VR devices. In *Proceedings of the 15th ACM SIGGRAPH Conference on Virtual-Reality Continuum and Its Applications in Industry-Volume 1*. ACM, 309–318.
[52] Shunichi Kasahara, Keina Konno, Richi Owaki, Tsubasa Nishi, Akiko Takeshita, Takayuki Ito, Shoko Kasuga, and Junichi Ushiba. 2017. Malleable embodiment: Changing sense of embodiment by spatial-temporal deformation of virtual human body. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 6438–6448.
[53] Grimes G. Kendall, A. and R. Cipolla. 2015. PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization.. In *Proceedings of the International Conference on Computer Vision (ICCV 2015)*.
[54] Konstantina Kilteni, Raphaela Groten, and Mel Slater. 2012. The sense of embodiment in virtual reality. *Presence: Teleoperators and Virtual Environments* 21, 4 (2012), 373–387.
[55] Elena Kokkinara and Mel Slater. 2014. Measuring the effects through time of the influence of visuomotor and visuotactile synchronous stimulation on a virtual body ownership illusion. *Perception* 43, 1 (2014), 43–58.
[56] Elena Kokkinara, Mel Slater, and Joan López-Moliner. 2015. The effects of visuomotor calibration to the perceived space and body, through embodiment in immersive virtual reality. *ACM Transactions on Applied Perception (TAP)* 13, 1 (2015), 3.
[57] Anya Kolesnichenko, Joshua McVeigh-Schultz, and Katherine Isbister. 2019. Understanding emerging design practices for avatar systems in the commercial social VR ecology. In *Proceedings of the 2019 on Designing Interactive Systems Conference*. 241–252.
[58] Lucas Kovar, Michael Gleicher, and Frédéric Pighin. 2008. Motion graphs. In *ACM SIGGRAPH 2008 classes*. ACM, 51.
[59] Huajun Liu, Xiaolin Wei, Jinxiang Chai, Inwoo Ha, and Taehyun Rhee. 2011. Realtime human motion control with a small number of inertial sensors. In *Symposium on interactive 3D graphics and games*. 133–140.
[60] Matthew R Longo, Friederike Schüür, Marjolein PM Kammers, Manos Tsakiris, and Patrick Haggard. 2008. What is embodiment? A psychometric approach. *Cognition* 107, 3 (2008), 978–998.
[61] Paterson M. H. Ma, Y. and E. Pollick. 2006. A motion capture library for the study of identity, gender, and emotion perception from biological motion. 38 (2006), 7291—-7299.
[62] Antonella Maselli and Mel Slater. 2013. The building blocks of the full body ownership illusion. *Frontiers in human neuroscience* 7 (2013), 83.
[63] Antonella Maselli and Mel Slater. 2014. Sliding perspectives: dissociating ownership from self-location during full body illusions in virtual reality. *Frontiers in human neuroscience* 8 (2014), 693.
[64] Alberto Menache. 2000. *Understanding motion capture for computer animation and video games*. Morgan kaufmann.
[65] Microsoft Corporation. 2010. Microsoft Kinect Games. Retrieved 2021 from https://en.wikipedia.org/wiki/Category:Kinect_games
[66] Kulpa R. Multon, F. and B. Bideau. 2008. Mkm: Aglobal framework for animating humans in virtual reality applications. *Presence: Teleoper. Virtual Environ.* 17 (2008), 17–28.
[67] Xuan Thanh Nguyen, Thi Duyen Ngo, and Thanh Ha Le. 2019. A Spatial-temporal 3D Human Pose Reconstruction Framework. *arXiv preprint arXiv:1901.02529* (2019).
[68] TechCrunch Oculus. 2020. TechCrunch - Oculus. Retrieved 2020 from https://techcrunch.com/2016/10/06/facebook-social-vr/
[69] Mathias Parger, Joerg H Mueller, Dieter Schmalstieg, and Markus Steinberger. 2018. Human upper-body inverse kinematics for increased embodiment in consumer-grade virtual reality. In *Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology*. ACM, 23.
[70] Tabitha C Peck and Mar Gonzalez-Franco. 2020. Avatar embodiment. a standardized questionnaire. *Frontiers in Virtual Reality* 1 (2020), 44.
[71] Ivan Poupyrev, Mark Billinghurst, Suzanne Weghorst, and Tadao Ichikawa. 1996. The go-go interaction technique: non-linear mapping for direct manipulation in VR. In *ACM Symposium on User Interface Software and Technology*. Citeseer, 79–80.
[72] Katja Rogers, Jana Funke, Julian Frommel, Sven Stamm, and Michael Weber. 2019. Exploring interaction fidelity in virtual reality: Object manipulation and whole-body movements. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–14.
[73] Enrique Rosales, Jafet Rodriguez, and Alla Sheffer. 2019. SurfaceBrush: from virtual reality drawings to manifold surfaces. *arXiv preprint arXiv:1904.12297* (2019).
[74] Charles Rose, Michael F Cohen, and Bobby Bodenheimer. 1998. Verbs and adverbs: Multidimensional motion interpolation. *IEEE Computer Graphics and Applications* 18, 5 (1998), 32–40.
[75] Philip Sedgwick. 2012. Pearson's correlation coefficient. *Bmj* 345 (2012).
[76] Park H. S. Sigal Y. Leonid abd Sheikh Shiratori, T. and K. J. Hodgins. 2011. Motion capture from body-mounted cameras.. In *SIGGRAPH 2011*. 7291—-7299.
[77] m. Slater. 2017. Implicit Learning Through Embodiment in Immersive Virtual Reality.. In *In D. Liu, D., Dede, C., Huang, R. and Richards, J. eds., Virtual, Augmented, and Mixed Realities in Education*. 19–34.

[78] Ronit Slyper, Guy Hoffman, and Ariel Shamir. 2015. Mirror puppeteering: Animating toy robots in front of a webcam. In *Proceedings of the Ninth International Conference on Tangible, Embedded, and Embodied Interaction*. ACM, 241–248.

[79] Bernhard Spanlang, Jean-Marie Normand, David Borland, Konstantina Kilteni, Elias Giannopoulos, Ausiàs Pomés, Mar González-Franco, Daniel Perez-Marcos, Jorge Arroyo-Palacios, Xavi Navarro Muncunill, et al. 2014. How to build an embodiment lab: achieving body representation illusions in virtual reality. *Frontiers in Robotics and AI* 1 (2014), 9.

[80] Zhipeng Tan, Yuning Hu, and Kun Xu. 2017. Virtual Reality Based Immersive Telepresence System for Remote Conversation and Collaboration. In *International Workshop on Next Generation Computer Animation Techniques*. Springer, 234–247.

[81] Manos Tsakiris and Patrick Haggard. 2005. The rubber hand illusion revisited: visuotactile integration and self-attribution. *Journal of Experimental Psychology: Human Perception and Performance* 31, 1 (2005), 80.

[82] Zhifu Wang, Xianfeng Yuan, and Chengjin Zhang. 2019. Design and Implementation of Humanoid Robot Behavior Imitation System Based on Skeleton Tracking. In *2019 Chinese Control And Decision Conference (CCDC)*. IEEE, 3541–3546.

[83] Johann Wentzel, Greg d'Eon, and Daniel Vogel. 2020. Improving Virtual Reality Ergonomics Through Reach-Bounded Non-Linear Input Amplification. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[84] Johann Wentzel, Greg d'Eon, and Daniel Vogel. 2020. Improving Virtual Reality Ergonomics through Reach-Bounded Non-Linear Input Amplification. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, Article 558, 12 pages. https://doi.org/10.1145/3313831.3376687

[85] Shihong Xia, Congyi Wang, Jinxiang Chai, and Jessica Hodgins. 2015. Realtime style transfer for unlabeled heterogeneous human motion. *ACM Transactions on Graphics (TOG)* 34, 4 (2015), 119.

[86] Jackie Yang, Christian Holz, Eyal Ofek, and Andrew D Wilson. 2019. Dreamwalker: Substituting real-world walking experiences with a virtual reality. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. 1093–1107.

[87] M Ersin Yumer and Niloy J Mitra. 2016. Spectral style transfer for human motion between independent actions. *ACM Transactions on Graphics (TOG)* 35, 4 (2016), 137.