HOOV: Hand Out-Of-View Tracking for Proprioceptive Interaction using Inertial Sensing

Paul Streli
Department of Computer Science
ETH Zürich, Switzerland

Rayan Armani
Department of Computer Science
ETH Zürich, Switzerland

Yi Fei Cheng
Department of Computer Science
ETH Zürich, Switzerland

Christian Holz
Department of Computer Science
ETH Zürich, Switzerland

Figure 1: *HOOV* is a wireless sensing method that complements existing virtual and augmented reality headsets to support hand tracking outside the field of view of the headset's cameras. In this example app of our system, a participant is building a structure composed of blocks with different sizes. When the user places an object onto the structure, his hand is tracked by the headset's cameras (*left*). As the user reaches for a block just to his right (*right*), his hand leaves the field of view of the cameras. Now, HOOV continuously estimates the 6D position and orientation of the wrist using the 6-axis inertial measuring unit attached to the wrist during the time the hand stays outside the field of view.

ABSTRACT

Current Virtual Reality systems are designed for interaction under visual control. Using built-in cameras, headsets track the user's hands or hand-held controllers while they are inside the field of view. Current systems thus ignore the user's interaction with *off-screen content*—virtual objects that the user could quickly access through proprioception without requiring laborious head motions to bring them into focus. In this paper, we present *HOOV*, a wrist-worn sensing method that allows VR users to interact with objects *outside* their field of view. Based on the signals of a single wrist-worn inertial sensor, HOOV continuously estimates the user's hand position in 3-space to complement the headset's tracking as the hands leave the tracking range. Our novel data-driven method predicts hand positions and trajectories from just the continuous estimation of hand orientation, which by itself is stable based solely on inertial observations. Our inertial sensing simultaneously detects finger pinching to register off-screen selection events, confirms them using a haptic actuator inside our wrist device, and thus allows users to select, grab, and drop virtual content. We compared HOOV's performance with a camera-based optical motion capture system in two folds. In the first evaluation, participants interacted based on tracking information from the motion capture system to assess the accuracy of their proprioceptive input, whereas in the second, they interacted based on HOOV's real-time estimations. We found that HOOV's target-agnostic estimations had a mean tracking error of 7.7 cm, which allowed participants to reliably access virtual objects around their body without first bringing them into focus. We demonstrate several applications that leverage the larger input space HOOV opens up for quick proprioceptive interaction, and conclude by discussing the potential of our technique.

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04…$15.00
https://doi.org/10.1145/3544548.3581468

CCS CONCEPTS

• **Human-centered computing** → **Virtual reality**; *Interaction devices*; • **Computing methodologies** → Neural networks; **Tracking**; Motion capture.

KEYWORDS

Virtual Reality, Hand Tracking, Inertial Sensing, Inertial Tracking, Proprioceptive Interaction, Eyes-free Interaction, Sensor Fusion

ACM Reference Format:

Paul Streli, Rayan Armani, Yi Fei Cheng, and Christian Holz. 2023. HOOV: Hand Out-Of-View Tracking for Proprioceptive Interaction using Inertial Sensing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3544548.3581468

1 INTRODUCTION

Mixed Reality systems are increasingly designed to be standalone, both for Augmented Reality (AR) as well as Virtual Reality (VR) scenarios. They typically integrate all necessary components for an immersive virtual experience in a user-worn headset. These include a stereoscopic display, stereo sound, processors for rendering, as well as sensors such as inertial measurement units (IMUs) and cameras for inside-out head-pose tracking in world space.

Standalone systems are highly portable, allowing them to operate in mobile scenarios, but requiring them to make several trade-offs in their design. Because their provided comfort strongly affects the user experience, size, form factor, and weight are the limiting factors in the implementation. These considerations ultimately limit the available computing power, battery capacity, and display size, which confines the *visible* field of view (FOV) for the wearer. They also affect the number, type, and placement of integrated sensors, in particular the cameras that track the user's surroundings and, thus, the effective *tracking* FOV. In addition to scene perception, these embedded cameras increasingly deliver the signal for detecting user input through hand gestures and interaction with virtual and physical objects. Therefore, the effective tracking FOV also determines the operational range for hand input on these devices.

In practice, today's headsets typically cover an operational range that slightly exceeds the user's visible FOV, thereby requiring all hand-object interaction to happen under *visual control*. While the operational range could be increased with additional headset cameras, which would incur additional needs for compute, power, and physical space inside the headset, certain areas around the body might still be outside the line of sight due to occlusion caused by clothing, hair, or other parts of the user's body.

As a result, today's systems implicitly require all interaction to occur in the visual FOV, which neglects most of the available space around the user for input. We argue that this is a missed opportunity, as humans perceive more than 210° [53], allowing us to perform hand-object interactions even at the edge of our periphery. Even without visual control, we routinely rely on proprioception to quickly place and retrieve physical objects in our vicinity [40, 64]—without requiring a turn of the head, which would slow down such motions. Since such short-term use cases do not justify the cost of additional cameras, *off-screen interaction* is not part of the interaction vocabulary on today's headsets. And with mixed reality devices' ever-decreasing form factor, we do not expect future devices to substantially widen the tracking FOV.

In this paper, we introduce *HOOV*, a method to track interaction outside a headset's tracking field-of-view in immersive environments. Using our novel data-driven inertial estimation pipeline, HOOV complements current headsets by leveraging the continuous signals from a 6-axis IMU inside a wrist-worn band to estimate the user's current hand position outside the headset's tracking FOV.

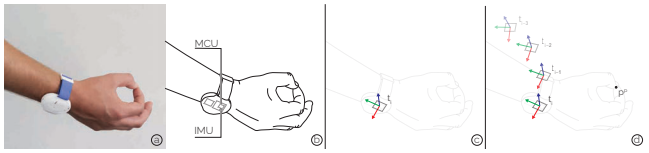

Figure 2: HOOV's tracking pipeline to predict outside field-of-view hand positions. (a) The wrist-worn band contains an embedded platform that houses (b) a 3-axis accelerometer and a 3-axis gyroscope. (c) From the signals of these inertial measurement units (IMU), HOOV first estimates the wrist's 3D orientation and (d) feeds the series of orientation estimates as input into our novel temporal machine learning model to predict the position of the hand.

1.1 Outside field-of-view interaction

Figure 1 shows an example application of our method. Here, a user is building a structure composed of blocks with different sizes, placed by his side for easy access. The user can pick up a building block by pinching it, dragging it to the tower, and dropping it at the desired location and with the hand-held orientation. The building blocks by his side replenish, such that grabbing one immediately produces another one at the same location.

Having built up the tower so far, the user has internalized the location of building blocks by his side by now. Naturally, as time progresses, he has to rely less and less on visual operation and, instead, simply reaches for one of the three locations to grab the corresponding piece. This aptly supports his construction process as the complexity of the structure advances, allowing him to keep his visual attention fixated on the tower and the spot where he plans to place the next block, while reaching out to grab it.

The underlying implementation of this is powered by our method HOOV. When the user's hand leaves the headset's tracking FOV, HOOV takes over the tracking and continuously provides the hand's current 3D position to the application. For this, HOOV processes the signals from the 3-axis accelerometer and 3-axis gyroscope sensor integrated inside the wrist-worn band as input. From these, HOOV estimates the wrist's current 3D orientation and feeds it into our novel temporal machine learning model, which estimates the current hand position.

To register input commands from the user, HOOV leverages the same IMU to detect hand gestures. In Figure 1, the user pinches to grab the intended building block, which HOOV recognizes as a performed gesture in the signal stream. HOOV's pinch detection complements the interaction paradigm on today's VR platforms, in which pinching is commonly used either for direct or remote selection through a ray [43, 45].

Performance Evaluation. We conducted two studies to evaluate HOOV's performance during outside field-of-view interaction in stationary scenarios. Both studies consisted of three tasks each. In the first task, participants grabbed objects outside their FOV while

facing forward, resembling our examples in Figure 1 but scaled up to 17 potential targets. In the second task, participants placed virtual spheres into 1 of 17 discrete drop zones by their side. In a final compound task, participants repeatedly switched between retrieving and placing objects outside their FOV, which constituted a compound and, thus, longer task. In Study 1, participants' wrists were tracked by an 8-camera OptiTrack Prime 13 system, permitting an analysis of the human ability to interact based on proprioception only, under nearly optimal tracking conditions. Participants achieved a success rate of ∼91% when grabbing and dropping objects off-screen. In Study 2, participants were operating exclusively using HOOV's tracking when their hand left the headset's view. They were successful in 86.18%, 87.65% and 86.18% of cases for the grab, placement, and compound task respectively.

When considered as a target-agnostic 3D tracking system, HOOV's simulated position estimates form a 3D trajectory that has a mean absolute error of 7.77 cm from the OptiTrack's reference path in Study 1. We observed that HOOV's error was lowest within the first three seconds, covering the duration of most interactions outside the user's field of view. In our simulations of lower tracking FOV in the headset (i.e., earlier moments at which the headset hands off tracking to HOOV), we found that HOOV's tracking error slightly increased, with a mean absolute error of 9.28 cm in a simulated 120° tracking FOV and a 10.16 cm error in a simulated 90° tracking FOV. In Study 2, the average error remained below 16 cm during real-time interactive use over longer off-screen trajectories. In its current implementation, HOOV's inference pipeline runs in real-time on a desktop machine with an *NVIDIA GeForce RTX 3090* GPU, requiring 4 ms to predict a 3D hand pose, which amounts to a maximum update frequency of 250 Hz.

In both studies, participants also completed a separate condition of the same tasks under visual control. In this condition, the hand was only tracked by the headset, in our case an Oculus Quest 2, requiring participants to turn their head to keep the hand inside the headset's field of view. While the average success rate rose to 95% across all tasks in Study 2, the average task completion time also increased, most notably by the compound task. Due to the monotony and repetitiveness of the task, most participants pointed out the unfavorable need for turning their heads in this condition and the strain this puts on their neck over time, preferring interaction through HOOV over interaction under visual control.

Applications. HOOV enables several scenarios that are directly applicable and useful for immersive systems. We show this at the example of three demo applications that benefit from the knowledge of the current hand position beyond the trackable FOV of the headset's cameras. In the first application, users can interact with building blocks outside their FOV as illustrated in Figure 1 and Figure 10. In the second scenario, HOOV improves a *Beat Saber* gaming experience by tracking quick arm movements that cross hardly tracked areas around the user's body. Finally, HOOV facilitates tracked arm movements of extended range as demonstrated in our third application, an archery game.

We conclude this paper with a discussion of our applications and the implications of our method for future immersive systems.

1.2 Contributions

In this paper, we contribute

- a tracking method that enables short-term interaction beyond the FOV of the cameras inside AR/VR headsets. Our method merely relies on the signals from a 6-axis IMU, integrated into our custom inertial sensing device placed on the user's wrist, which captures the information for outside-field-of-view 6D hand tracking and pinch detection and provides haptic feedback during interaction events.
- a novel temporal deep learning architecture that comprises a Transformer for the estimation of the current hand position from the IMU signals and the latest available output of the headset's visual hand tracking pipeline.
- a user study with 12 participants to investigate human proprioceptive abilities and HOOV's tracking performance, where participants retrieved and placed spheres at 17 different target locations outside the visual and the camera's field of view. Using an 8-camera OptiTrack system, participants grabbed and placed the correct object ∼91% of the time. In an offline simulation, they would have grabbed and placed the correct object in ∼85% of all cases using HOOV. Compared to a baseline with a commodity headset where participants grabbed and placed objects under visual control, their success rate increased to ∼94% at the expense of 20% slower task completion and repeated neck movement.
- a user study with 10 participants where participants completed the selection and placement tasks using HOOV in real-time, achieving success rates of more than 86%.

2 RELATED WORK

HOOV is related to body capture using worn sensors, inertial tracking, and interactions driven by spatial memory, proprioception, and kinesthesia.

2.1 Body capture using worn sensors

Information about the user's body pose, especially the posture of the arm, enables VR and AR systems to correctly embody the user and to detect input through gestures.

In contrast to external body pose capture systems that include high-end commercial marker-based camera systems [42], depth and RGB [10, 16, 39] cameras as well as non-optical systems [26, 67, 68], body-worn systems enable tracking in mobile settings, and do not require the user to remain within a constraint area. To receive a wider view of the user's body for tracking, cameras with wide angle-lenses were attached to various parts of the body including the wrist [62], the feet [7], and the chest [25, 27] or alternatively integrated into controllers [3] or using a suspension on the headset [1, 46] further away from the user's body. However, these solutions tend to suffer from occlusion and are often obtrusive to wear. Alternative approaches directly estimate the full-body pose based solely on the available temporal motion information of the user's head and hands [2, 28, 61]. Moreover, various specialized mobile hand-held or body-worn systems have been built for tracking that make use of sensing modalities such as magnetic [11], mechanical [41], or acoustic [29, 57] sensing.

Prior research has also shown that body-attached IMUs can be used in a standalone system for tracking human pose. Depending

on the number of body locations that are instrumented, the problem varies in difficulty. Sparse Inertial Poser [59] estimates the 3D human pose based on 6 IMU sensors attached to the wrists, lower legs, the back and the head using a joint optimization framework incorporating anthropometric constraints. Deep Inertial poser [23] trains a recurrent neural network (RNN) for this task. Transpose [66] and PIP [65] further improve on the predicted output by incorporating joint-specific loss terms and physic-aware motion optimizers.

2.2 IMU-based arm and wrist tracking

Instead of tracking the full body, Li et al. attached three IMU sensors to the hand, forearm and upperarm to track a user's arm pose through solving inverse kinematics [35]. Closely related to our work, Shen et al. presented an online and an offline tracking method to estimate the posture of the arm based on a single wrist-worn 9-axis IMU integrated within a smartwatch [50]. The method finds an optimal path through a set of potential candidate arm postures that are retrieved from a dictionary based on the IMU-estimated wrist orientation at each time step. The offline version uses Viterbi decoding to estimate an optimal tracking path, and achieves a medium tracking error of 9.2 cm for the wrist. However, since it occurs a complexity of $O(N^3T)$, it is not suitable for real time-tracking. A simpler online version that estimates the arm position from the candidate postures at a single step using a frequency-weighted average achieves a median tracking error of 13.3 cm. Liu et al. [36] propose to improve the computation based on Hidden Markov Model state reorganization, achieving an accuracy of 12.94 cm. With an updated algorithm including a Particle Filter, a median error of 8.8 cm was reported in a static setting [49].

Similarly, Wei et al. propose an RNN architecture to track the arm posture from a 6-axis IMU assuming a fixed shoulder position [60]. They report a median error of 15.4 cm and a MAE of 16.4 cm for the wrist in a leave-one-subject-out evaluation scheme. LimbMotion uses an additional edge device for acoustic ranging to estimate the arm posture with a median wrist error of 8.9 cm [69].

Compared to the previous approaches, our method does not aim to support stand-alone arm tracking based on a 6-axis IMU but to support a visual tracking system for short periods of time where the hand moves outside the cameras field of view. This allows us to correct for drift, especially in the yaw direction which is more pronounced due to the missing magnetometer, as soon as the hand is visible by the headset's cameras. Moreover, our method does not assume a fixed shoulder position.

2.3 Spatial Memory, Proprioception, and Kinesthesia

The human ability to operate without visual control is mainly supported by the following factors: spatial memory [48], proprioception [13], and kinesthesia [34]. Spatial memory refers to the part of memory responsible for recording the position and spatial relations between objects [5, 17]. Proprioception refers to the sense of position and orientation of one's body parts with respect to each other [13, 51]. Kinesthesia, which is often used interchangeably with proprioception, refers to the perception of one's body movements and motions [34]. In prior research, there is substantial work focused on characterizing the aforementioned cognitive and perceptual abilities of people [8]. Gutwin et al. [19] and Cockburn and McKenzie [12], for instance, studied people's capabilities of building spatial memory in 2D and 3D spaces. Hocherman [22], Soechting and Flanders [52], and Medendrop et al. [37] examined the extent to which people can rely on their proprioception to perform target selections. Andrade and Meudell [4] and Postma and De Haan [44] showed that people are capable of learning spatial locations without paying particular attention to them.

Within the human-computer interaction community, it is generally acknowledged that spatial memory, proprioception, and kinesthesia can be exploited to support rich and efficient interactions [40, 48]. Li et al.'s Virtual Shelves technique [33] leveraged users' kinesthetic memory for triggering programmable shortcuts. Using Virtual Shelves, users can select shortcuts by pointing a spatially-aware device at 28 different locations in front of themselves. Yan et al. [64] build on top of this work and experimentally studied eye-free target acquisition in VR including the additional aspect of user comfort. Cockburn et al. [13] presented a design space exploration for the interaction termed "air pointing". Gustafson et al.'s Imaginary Interfaces [17] sought to enable bi-manual empty-handed interactions without visual feedback by relying on users' short-term memory. Imaginary phone [18] demonstrated the potential of transferring users' spatial memory from a familiar physical device to operating an "imaginary" equivalence on their palm. Additional works explored proprioception-driven interactions for menu usage [55], allowing information access via a hip-attached touch device [14], supporting mobile phone access for visually impaired users [34], and enabling interaction with the back of a headset [15].

Also closely related to HOOV, many works support users in proprioceptively performing interactions through haptics. Kim et al. [31] guided visually-impaired users in target selection on a large wall-mounted display with vibrotactile feedback. Barham et al.'s CAVIAR device [6] similarly used vibrotactile actuators to guide users' hands with continuous stimuli. Vo and Brewster [58] studied ultrasonic haptic feedback to support spatial localization.

HOOV ultimately aims to enable the aforementioned spatial memory- and proprioception-driven interactions by expanding the tracking space of current VR devices for hand-object interactions. We focus on addressing the technical challenge of performing the sort of eye-free interactions explored by Yan et al. [64] when the available headset tracking field-of-view is limited. We further support users in performing eye-free interactions with haptic feedback via an actuator integrated into our wrist-worn device.

3 HOOV METHOD: ESTIMATING PINCH POSITIONS FROM OBSERVATIONS OF WRIST ORIENTATIONS

We now introduce our method that enables the tracking of the wrist position, $\mathbf{p}^w \in \mathbb{R}^3$, and orientation, $\mathbf{R}^w \in SO(3)$, based on the captured acceleration, $\mathbf{a}^w \in \mathbb{R}^3$, and angular velocity signals, $\omega^w \in \mathbb{R}^3$, of a 6-axis IMU placed at the user's wrist. Besides the wrist-worn band's motion signals, HOOV also receives as input the headset's current position, $\mathbf{p}^h \in \mathbb{R}^3$, and orientation, $\mathbf{R}^h \in SO(3)$, as well as the sequence of τ_t last available head and hand poses $\{\mathbf{p}^w, \mathbf{R}^w\, \mathbf{p}^h, \mathbf{R}^h\}_{-(\tau_t-1):0}$ before the hand has left the tracking FOV

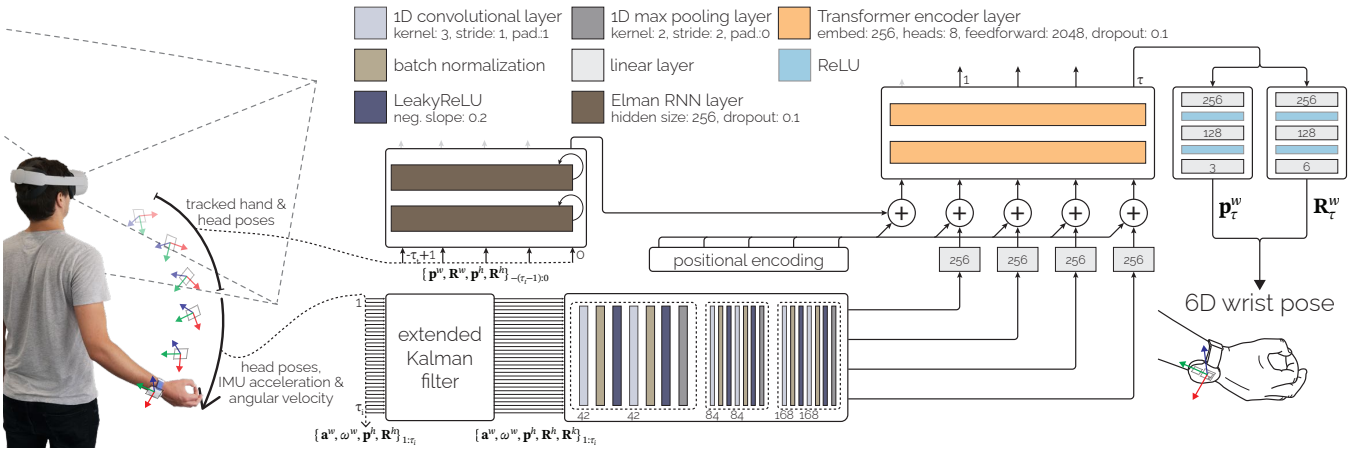

Figure 3: HOOV's wrist pose estimation pipeline. A 2-layer RNN receives the last 5 available hand poses from the Oculus Quest 2 as input. The inertial motion data captured since the point where the hand left the FOV of the headset's cameras is downsampled by a factor of 8. The inertial data is then transformed by a linear embedding layer and appended to the output of the RNN. A Transformer processes the sequence before each sample is mapped to a position and an orientation.

of the headset. We describe this mapping f by

$$\{\mathbf{p}^w, \mathbf{R}^w\}_{1:\tau} = f(\{\mathbf{a}^w, \omega^w, \mathbf{p}^h, \mathbf{R}^h\}_{1:\tau_i}, \{\mathbf{p}^w, \mathbf{R}^w, \mathbf{p}^h, \mathbf{R}^h\}_{-(\tau_t-1):0}),$$

where τ matches the number of considered time steps predicted since the hand has left the tracking FOV, and τ_i is the number of inputs sampled in the same period.

This is a challenging problem since we only receive the noisy observations of the relative motions of the wrist whose position lies in a 5-DoF space given a static shoulder position (3 DoF through shoulder joint and 2 DoF through lower arm) [49] and a 7-DoF space given a constant neck position (additional 2 DoF from clavicle) as input. However, the set of possible wrist poses is severely constrained given the knowledge of the forearm's orientation and by incorporating the anatomical constraints of the individual human joints [49, 50]. While orientation estimation from a 6-axis IMU suffers from yaw drift due to the lack of a magnetometer, we demonstrate that a learning-based method can correct for this drift for short-term out-of-view interactions in static settings. Our method does this by estimating the most likely arm trajectory from an implicitly acquired approximated distribution of previously seen trajectory observations.

In addition to tracking the 3D positions of the wrist, HOOV detects input commands from the wearer. We continuously process the signals of the accelerometer for characteristic patterns to detect gestures, such as pinching or fist clenching. Using the build-in haptic actuator, our prototype is capable of rendering haptic feedback to the user in response via a knocking sensation.

3.1 Inertial 6D hand pose estimation

Figure 3 shows an overview of our inertial 6D hand pose estimation pipeline.

Input and output representation. The input consists of the accelerometer and gyroscope signals from the IMU within the wrist-worn band as well as the current head position and orientation estimated by the inside-out tracking of the VR headset. The spacing between the samples of tracked head positions is matched to the sampling interval of the IMU through cubic interpolation. We further estimate an initial orientation of the wrist by directly applying an extended Kalman Filter [30] to the gyroscope and accelerometer signals [20, 47]. We convert all orientations to their 6D representation to ensure continuity [71]. For each time step, we concatenate the acceleration, angular velocity, head position and orientation, and the output of the extended Kalman Filter $\mathbf{R}^k$. This input sequence, $\mathbf{X} \in \mathbb{R}^{\tau_i \times 21}$, starts from the moment when the hand leaves the trackable FOV of the headset. We obtain another input sequence $\mathbf{S} \in \mathbb{R}^{\tau_t \times 18}$ of the last τ_t head and hand poses that were tracked by the headset before the hand left the tracking FOV.

Based on this input, our estimator directly predicts wrist position $\mathbf{p}^w \in \mathbb{R}^3$ and rotation $\mathbf{R}^w \in \mathbb{R}^6$ within the world. Using the orientation of the wrist, we produce a more refined estimate of the pinch position, $\mathbf{p}^p \in \mathbb{R}^3$, by applying an offset of 15 cm to the wrist position,

$$\mathbf{p}^p = \tilde{\mathbf{R}}^w(0, -0.15, 0)^T + \mathbf{p}^w,$$

where $\tilde{\mathbf{R}}^w$ is the rotation matrix corresponding to $\mathbf{R}^w$ and the y-axis of the right-handed local coordinate system centered at the wrist points along the forearm to the shoulder with the z-axis pointing downwards through the palm. We took this option because we expected position estimates to be better than representations that encode the rotations for each joint along the kinematic tree from the head to the wrist. Since the interaction happens outside the user's visible FOV, visualization artifacts due to varying bone lengths are of limited concern.

Network architecture. HOOV relies on a neural network to approximate the mapping f. The architecture of the network consists of a downsampling module that receives the input sequence $\mathbf{X}$ containing the inertial motion information as input and reduces the sequence along the temporal axis by a factor of 8. The module

consists of three blocks that each reduce the signal's temporal resolution by a factor of 2 and consist of two convolutional layers with a kernel of size 3 followed by a max pooling layer. The samples of the remaining features are converted to linear embeddings that are fed to a Transformer consisting of two encoder layers. At the start of the sequence, we add an initial token that embeds the information from the headset-tracked hand and head poses $\mathbf{S}$, extracted through a 2-layer Elman RNN. We apply a sinusoidal positional encoding to the input of the Transformer [56], and avoid that future samples are attended to for the estimation of any output samples by applying a corresponding mask to the sequence.

For each output sample of the Transformer corresponding to the down-sampled features from the inertial input, we predict the position and orientation of the wrist using a series of fully-connected layers. This design supports sequences of variable length.

Loss function. Our loss function,

$$\mathcal{L} = \mathcal{L}_p + \mathcal{L}_R = \sum_{t=1}^{\tau} (|\hat{\mathbf{p}}_t^w - \mathbf{p}_t^w| + |\hat{\mathbf{R}}_t^w - \mathbf{R}_t^w|),$$

consists of two terms, where $\mathcal{L}_p$ penalizes the positional and $\mathcal{L}_R$ the rotational offset using the L1 loss function, and $\mathbf{p}_t^w$ and $\mathbf{R}_t^w$ are the ground-truth and $\hat{\mathbf{p}}_t^w$ and $\hat{\mathbf{R}}_t^w$ the predicted position and orientation of the wrist at time t respectively.

3.2 Pinch detection

To detect pinch events, we threshold the running rate-of-change score c, which accumulates the absolute change in the acceleration signals $\mathbf{a}^w$ captured by the IMU at the wrist across time [38, 54],

$$c_t = \tfrac{1}{D} c_{t-1} + |\|\mathbf{a}_t^w\|_2 - \|\mathbf{a}_{t-1}^w\|_2|.$$

Pinch events cause sudden changes in a_t^w that lead to a strong increase c_t, and thus, can be detected through a threshold. The exponential reduction factor D attenuates past accumulations.

4 IMPLEMENTATION

HOOV is implemented to run as a real-time interactive tracking system on an 8-core Intel Core i7-9700K CPU at 3.60 GHz with an NVIDIA GeForce RTX 3090 GPU. HOOV consists of three components: 1) a wrist-worn sensing platform, 2) a virtual reality interface, and 3) a central control and sensor fusion unit that handles the communication to the wristband hardware and the virtual reality headset as well as the estimation of the hand pose outside the FOV.

4.1 Hardware

We custom-designed an embedded platform for HOOV for sensing and actuation during interaction in mid-air. As shown in Figure 4, our electronics platform centers around a System-on-a-Chip (NRF52840, Nordic Semiconductors) that samples a 6-axis IMU (LSM6DSOX, STMircroelectronics) at 427 Hz. Our prototype streams the inertial data to a PC, either wirelessly over BLE or through a wired serial interface. The prototype integrates a separate board that embeds an audio amplifier (MAX98357A, Maxim Integrated) to drive a Lofelt L5 actuator to produce haptic feedback. All components of our prototype, including a battery, are housed in a 3D-printed case, which attaches to the wrist using a strap.

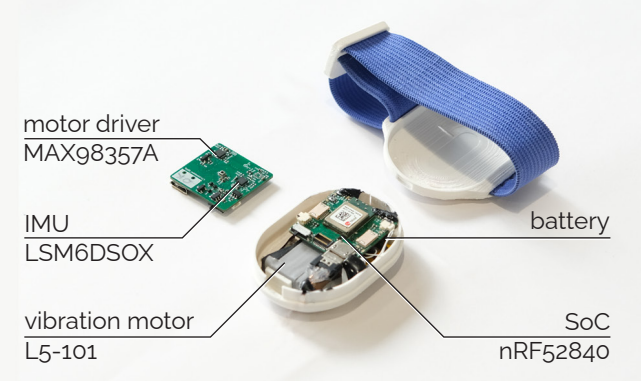

Figure 4: HOOV's wristband integrates an electronics platform with a 6-axis IMU and a System-on-a-Chip within a 3D-printed case. The case also contains a battery and a vibration motor to provide haptic feedback to the user.

4.2 Virtual Reality

We implemented our VR application in Unity 2021 for the Oculus Quest 2 VR headset. Our application communicates with our outside-field-of-view tracking pipeline via a web socket. As tracking source, our VR environment receives the hand pose computed by the headset, sampled at $\sim$ 70 Hz when the user's hand is tracked, and the estimates produced by HOOV otherwise. Input events are triggered through HOOV's pinch detection algorithm in both states.

4.3 Sensor fusion and outside FOV tracking

The main control unit of our implementation is a state machine that moves between states depending on the current tracking status of the VR headset. During the outside FOV state, the headset forwards the received head poses and the inertial input data together with the last 5 available hand and head poses from the VR headset to the deep learning-based hand pose estimation pipeline.

We implemented the main control unit in Python 3.8. The program runs across multiple cores to handle the communication with the sensing hardware and the processing of the corresponding signals, and to set the current hand pose for the VR interface.

Upon detecting a pinch gesture using an exponential reduction factor D of 1.07 and a threshold of 4.9 $\mathrm{m/s^2}$, our system triggers an event inside the VR environment and sends a command to the motor driver to activate the haptic feedback once.

HOOV's network is implemented in *PyTorch* and has 4,408,199 trainable parameters. We use the Kaiming initialization [21], the Adam optimizer [32] with a learning rate of 10^{-4}, and a batch size of 16 where we randomly group sequences of equal length.

5 STUDY 1: EVALUATION WHERE PARTICIPANTS INTERACTED BASED ON OPTITRACK TRACKING

To evaluate HOOV's potential, we conducted a user study in which participants placed and retrieved objects outside their FOV. Study 1 served two purposes: (1) It allowed us to quantify HOOV's accuracy of tracking wrist positions outside the headset's tracking FOV. (2)

We could evaluate how quickly and accurately participants selected and placed objects outside their field of view compared to performing the same task under visual control by turning their heads. This quantified the effect of operating under proprioceptive control. To evaluate the upper bound of our approach, we conducted this first evaluation while participants interacted based on OptiTrack tracking to exclude the impact of tracking errors on participants' behavior. (See Section 6 for our real-time evaluation of HOOV for the same tasks (Study 2).)

5.1 Study design

5.1.1 Apparatus. Our experiment apparatus consisted of an Oculus Quest 2 VR headset, a HOOV wristband as described above, a computer for logging, and an optical motion capture system.

VR headset. Throughout the study, participants wore a Quest 2 VR headset with integrated visual hand tracking. For the study setup, we implemented a VR environment featuring objects, placement zones, and task protocols in Unity 2021. Participants saw a visualization of their right hand when it was within their FOV.

HOOV's IMU streams and haptic feedback. Participants wore a HOOV wristband on their right arm, which continuously streamed the data from the IMU to a PC for processing. From the continuous stream of accelerations, our apparatus detected pinch events as described in Section 3.2. To eliminate the impact of wireless connectivity onto participants' performance, we connected the wristband through thin and flexible magnet wires to the PC for power and reliable serial communication in this study.

Motion capture for 3D wrist and head trajectories. To compare the accuracy of HOOV's estimated wrist position to a ground-truth baseline, we tracked rigid-body markers attached to the participant's headset, shoulder, elbow, and wristband using an 8-camera OptiTrack system with sub-mm accuracy to obtain positions and rotations at a sampling frequency of 240 Hz. The rigid bodies are made out of cardboard that we attached to worn elbow and shoulder pads, and the HOOV wristband (Figure 5). To track the headset, we directly attached four 9-mm tracking markers to the device.

An experimenter ensured a consistent placement of the rigid bodies across participants ahead of the experiment and verified throughout that the markers did not slip. We calibrated the OptiTrack to the coordinate system of the VR headset through two pre-defined calibration points in the physical space that we marked with a controller in the virtual environment.

5.1.2 Participants. We recruited 12 participants from our local university (3 female, 9 male, ages 22–35, mean=25.7 years). Using the OptiTrack, we measured the distance between the worn headset and the floor as well as the distance between the neck and the wrist while participants stood in a T-pose. The average headset distance from the ground was 173 cm (SD=12 cm, min=149 cm, max=188 cm), and arm lengths ranged between 62 cm and 80 cm (mean=72 cm, SD=6 cm). Participants self-reported their prior experience with VR technology on a 5-point Likert scale (from 1–never to 5–more than 20 times). Participants' ratings ranged between 1 and 5 and their median prior experience was 3. Each participant received a small gratuity for their time.

5.1.3 Task. The study consisted of three grab-and-place tasks. Throughout the experiment, participants stood at a fixed point that was highlighted on the floor using tape, such that the experimenter could verify their position.

In the first task GRAB (Figure 5), participants stood next to a set of 17 spherical target objects outside their FOV in the virtual environment, grabbed an intended sphere, and placed it into the blue dropzone in front of them. Participants saw a down-sized illustration of the task environment in front of them, which highlighted the sphere to pick next. Participants grabbed a sphere using a pinch gesture, selecting the sphere that was closest to their right hand when in range. Participants then moved it towards the dropzone and released it using a second pinch.

In the second task DROP, participants picked up a spherical target in front of them, and placed it within one of 17 drop zones to their right outside their FOV. Again, a set of three matching illustrations highlighted the task's target drop zone next to them. Participants first grabbed the sphere using a pinch gesture, before moving it to the target drop zone. After pinching again, the apparatus placed the sphere in the drop zone that was closest to the hand's position at the time of the second pinch.

The third task COMPOUND combined Tasks 2 and 1. First, participants grabbed the sphere within their field of view in front of them and dropped it into one of the 17 drop zones outside their field of view as instructed (Figure 5b). Then, they were instructed to grab one of the 17 spheres from outside their field of view and release it within the drop zone in front of them (Figure 5a).

As shown in Figure 5, the spheres and drop zones were arranged in a grid of 3 rows, 3 columns, and 2 layers in the lateral direction of the participant. We excluded the closer center location behind the participant, because reaching it would require contorting one's arm in an uncomfortable pose without turning one's upper body. All other targets outside the FOV were convenient to reach.

The spacing of the targets and drop zones remained static throughout the study, and was adjusted at the beginning of the experiment to each participant's eye level and arm length to ensure all targets could be reached. The spacing in the lateral direction between the two layers was equal to half the length of the participant's arm. The spacing between objects and drop zones in the sagittal plane was equal to half the distance between the neck and the wrist of the participant when performing a T-pose, corresponding to an angular spacing of around 30° between the drop zones and targets further away from the participant. The target and drop zone grid was centered at the height of the participant's shoulder, and placed at a distance so that the participant's hand would reach all elements of the second layer when elongated.

5.1.4 Conditions. Participants performed the three tasks in two conditions.

In the OCULUS condition, the participant's hand was solely tracked by the Quest 2 headset. Thus, participants needed to turn their heads during the task and follow their hands to ensure that the headset tracked the hand's position when performing the task.

In the OptiTrack condition, participants were instructed not to turn their head and to look straight forward. An experimenter ensured this throughout by monitoring the head-mounted display

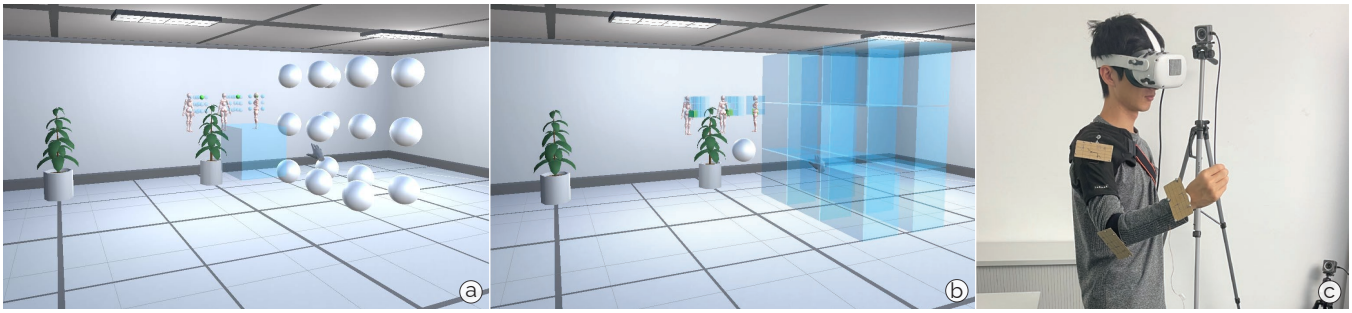

Figure 5: Study environment. Participants stood next to the set of spheres, focusing their gaze towards the wall onto a highlighted position, such that the spheres were to their right outside the headset's field of view. (a) In the first task GRAB, participants grabbed one of the 17 spheres outside their FOV and placed it in front of them into the blue cube. (b) In the second task DROP, participants grabbed the sphere in front and dropped it into one of the 17 drop zones outside their FOV. In the third task COMPOUND, participants dropped a sphere into a drop zone outside their FOV (b), and then retrieved another sphere outside their FOV (a). (c) Participants wore a Quest 2 headset and our apparatus tracked their shoulders, elbows, and wrists using rigid-body markers and a high-resolution OptiTrack system.

through a secondary screen. During this condition, the wrist position and orientation was provided to the VR environment by the OptiTrack motion capture system.

In both conditions, the apparatus detected pinch gestures for grabbing and releasing the spheres using the IMU inside the HOOV wristband as in Section 3.2. Participants received haptic feedback for detected pinch events and were notified of erroneous selections and placements with audio feedback.

5.1.5 Procedure. The study started with a brief introduction of the tasks. The experimenter then noted down the participant's age and gender. Participants performed a T-pose to calibrate the system to their body size. The study started with a training session of all conditions and tasks, followed by the experiment.

During training, participants performed 17 trials, one for each target location, for each task. When using the OptiTrack condition, participants received visual guidance for the DROP and GRAB task through the down-sized visualization to allow them to refine their internal model of their hand position in 3-space. The visualization highlighted the corresponding sphere or drop zone as soon as the participant's hand came into contact with it. This online highlighting was exclusive to the training session and was not available during the actual experiment.

After training, participants performed all tasks in each of the conditions without guidance. We counterbalanced the order of the task and conditions across participants. Participants performed 34 trials for DROP and GRAB, and 68 trials for the COMPOUND task, where a single trial consisted of placing or retrieving a single target object. For DROP and GRAB, each drop zone and sphere was used as target twice with the order across targets randomized. The same drop zone or target sphere was not used in direct succession. For the COMPOUND task, we fixed the order of locations between directly successive DROP and GRAB subtasks but randomized the order across combinations. We ensured that each valid position was used twice as drop zone and target object location. In total, participants completed 2 conditions × (34 DROP trials + 34 GRAB + 68 COMPOUND trials) = 272 trials in under 45 minutes. Participants were instructed to complete the tasks as fast as possible while avoiding mistakes.

5.1.6 Measures. As dependent variables, our apparatus measured the time it took participants to complete each trial and the percentage of successfully completed trials. A trial was successful when the participant placed the correct sphere in the correct drop zone.

5.2 Comparing HOOV with high-accuracy tracking from a motion capture system

Our study apparatus logged all signals and, thus, recorded a labeled dataset of IMU signals and ground-truth wrist and head positions from the motion capture system.

We used the recordings to train and test HOOV's estimation network for an offline evaluation. This allowed us to not only quantify HOOV's performance given the specific configuration of the headset, but we could also simulate a variety of FOV configurations and, thus, moments at which HOOV took over tracking from the headset due to our knowledge of the hand positions relative to the head pose at every time step. The OptiTrack condition thereby acts as the upper bound for participants' overall performance given the tracking system's high accuracy. Using the OptiTrack data, we could evaluate participants' ability to interact with objects outside their FOV under near-perfect tracking conditions. This enables us to decompose HOOV's error into contributions from participants' inherent limitation to perform proprioceptive interactions and limitations associated with our method itself.

Training and evaluation. We evaluated our network in a 6-fold cross-validation where each fold used the data of two randomly selected participants for testing and the other 10 participants for training. We further augmented our training data with the recordings from three additional participants that we acquired while piloting our study setup. For training, we use all data from the OptiTrack condition, including the trajectories recorded during participants' initial training sessions. To further augment our dataset, we trained our network on sequences that simulate a horizontal FOV between

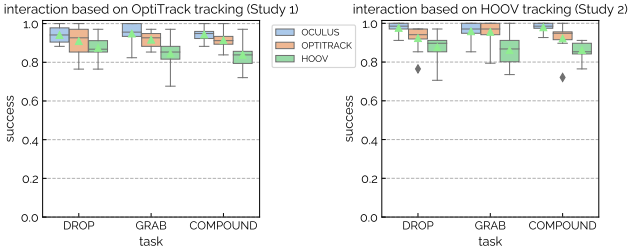

Figure 6: Mean success rates during the tasks for Study 1 and Study 2. (Left) In Study 1, participants were most accurate when bringing targets into view (OCULUS condition), with minor error rates when interacting outside their field of view (OptiTrack condition) due to proprioceptive capabilities. With HOOV, participants reached 93% to 96% of this accuracy. (Right) Error rates remained comparable during our online evaluation in Study 2 where HOOV was predicting input locations in real-time.

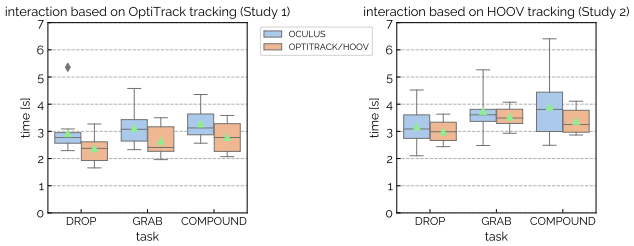

Figure 7: Mean trial duration of task completion for Study 1 and Study 2. (Left) On average, mean completion time was longer in all three conditions when operating under visual control in Study 1. (Right) In Study 2 with a fresh set of participants who were inexperienced VR users, mean completion time in the OCULUS condition was again longer, most pronounced in the COMPOUND task.

40° and 120° in steps of 5°. We also generated training input based on the logged tracking information from the Quest 2.

We trained our network until convergence on validation data from a participant that was extracted and excluded from each training dataset. We trained for at least 250,0000 iterations on an NVIDIA GeForce RTX 3090, which takes approximately 12 hours. For the test set, we only use the trials of the OptiTrack condition from the actual experiment.

We simulated and evaluated three headset tracking FOVs, including 90° (i.e., the visible FOV of the Quest 2), 120° (i.e., the approximate FOV of human binocular vision), and the actual tracking FOV of the Quest 2. The latter exceeds 120° as we empirically determined through the headset's reported tracking state.

We evaluated HOOV's output in terms of mean absolute distance error (MAE) and mean angle of the difference rotation (MAD) [24], $\theta = 2 \arccos(|\hat{\mathbf{q}}^w \cdot \mathbf{q}^w|)$, to the measurements reported by the motion capture system, averaged over the whole out-of-field-of-view tracking path. Here, $\hat{\mathbf{q}}^w$ and $\mathbf{q}^w$ are the unit quaternions corresponding to the ground-truth orientation $\hat{\mathbf{R}}^w$ and predicted orientation $\mathbf{R}^w$ respectively, and $\cdot$ denotes the inner (or dot) product of vectors. We also compare HOOV's success rates for the simulated FOV ranges to OCULUS and OptiTrack.

5.3 Results

5.3.1 Success rate. Figure 6 (left) shows an overview of participants' performance success during this study. Using OCULUS, participants were careful not to make mistakes when performing the tasks under visual control. Participants reached an average success rate of 94.12% for DROP (max=100.00, min=88.24, SE=1.30), 94.61% for GRAB (max=100.00, min=79.41, SE=1.69), and 94.12% for COMPOUND (max=100.00, min=88.24, SE=0.93). Using OptiTrack, participants' mean success rate was 90.93% for DROP (max=100.00, min=76.47, SE=1.91), 90.69% for GRAB (max=97.06, min=82.35, SE=1.50), and 90.56% for COMPOUND (max=100.00, min=83.82, SE=1.26).

When simulating the tracking of participants' wrist motions outside the FOV with HOOV, participants' overall success rate was 87.50% for DROP (SE=1.63, max=97.06, min=76.47), 84.56% for GRAB (SE=2.02, max=97.06, min=67.65) and 83.82% for the COMPOUND task (SE=1.83, max=97.06, min=72.06).

5.3.2 Trial Duration. As shown in Figure 7 (left), participants' mean completion time was shorter for the three tasks when not looking at the outside FOV regions, i.e., during GRAB, DROP, and COMPOUND in the OptiTrack condition. On average, participants took 2.37 s to complete a DROP trial (max=3.27, min=1.66, SE=0.14), 2.65 s for a GRAB trial (max=3.50, min=1.96, SE=0.15), and 2.79 s for a COMPOUND trial (max=3.59, min=2.07, SE=0.17). As shown in Figure 7, participants' mean completion time increased across all tasks in the OCULUS condition. Here, the average mean trial completion time was 2.93 s for DROP (max=5.36, min=2.29, SE=0.22), 3.12 s for GRAB (max=4.58, min=2.33, SE=0.17), and 3.29 s for the COMPOUND task (max=4.36, min=2.57, SE=0.16).

5.3.3 Absolute tracking accuracy of HOOV. Comparing the wrist poses estimated by HOOV to the poses captured by the OptiTrack system when participants' hands were outside the tracking FOV of the Quest 2, we found an MAE of 7.77 cm across participants. The median position offset averaged across participants was 7.00 cm. Figure 8 illustrates the development of the average tracking error over time from the point when a participant's hand left the FOV of the headset.

When dropping an object outside the FOV, HOOV's position estimates had an average MAE to the actual hand position of 6.95 cm. The average MAE when selecting objects outside the FOV was 7.72 cm. The orientation error as measured by the mean angle of the difference rotation is 6.50° (see Table 1).

As mentioned above, we simulated a series of tracking FOV to assess how HOOV's tracking capabilities may deteriorate for larger areas outside the FOV and, thus, increased of HOOV-based tracking. For a simulated horizontal FOV of 120°, the average positional MAE across participants increased to 9.28 cm, and the average MAD increased to 7.67°. The offset to the actual release position and grab position while performing the DROP and GRAB stayed within 10 cm, amounting to 8.92 cm and 9.86 cm, respectively, on average.

Table 1: The table shows the tracking error of HOOV with headsets of three different horizontal FOV ranges compared to a high-end motion capture system. The values are averaged across all participants. (±σ) is the standard deviation across participants. *mean pos.* is the mean absolute error (MAE), *median pos.* the median error and *std. pos.* is the standard deviation within participants in terms of position in cm. *mean rot.* is the mean angle of the difference rotation (MAD) and *median rot.* the median angle of the difference rotation in °. *std. rot.* is the standard deviation of the angle of the difference rotation within participants. *mean DROP*, *mean GRAB*, and *mean COMP.* are the mean distances between the actual and HOOV's estimate hand position at out-of-field-of-view release and grab events for the DROP, GRAB, and COMPOUND task, respectively.

FOV	mean pos.	median pos.	std. pos.	mean DROP	mean GRAB	mean COMP.	mean rot.	median rot.	std. rot.
> 120°	7.77 (±1.48)	7.00 (±1.55)	2.46 (±0.47)	6.95 (±2.90)	7.72 (±2.90)	8.21 (±2.95)	6.50 (±1.54)	5.41 (±1.30)	2.40 (±0.53)
120°	9.28 (±2.20)	8.56 (±2.45)	3.26 (±0.81)	8.92 (±3.64)	9.86 (±3.95)	10.35 (±4.00)	7.67 (±1.77)	6.49 (±1.54)	3.00 (±0.66)
90°	10.16 (±2.13)	9.53 (±2.28)	3.88 (±0.91)	10.20 (±4.08)	10.90 (±4.10)	11.23 (±4.17)	8.12 (±1.93)	7.02 (±1.54)	3.41 (±0.90)

For a simulated horizontal FOV of only 90°, HOOV estimated positions with an average MAE of 10.16 cm and estimated orientations with an MAD of 8.12°.

5.4 Discussion

The results of Study 1 underline the potential of our method HOOV to complement camera-based headset tracking for motions and interactions outside their tracking FOV. Most obviously was participants' speed increase when they could directly interact with targets outside their FOV without turning their heads to first bring them into the view. On average, mean completion times were 19% lower for the DROP and 15% lower for the GRAB and COMPOUND task when participants were able to interact outside their FOV. Because participants were able to directly compare the conditions during their experience in the study, several mentioned, unprompted, that turning their heads felt like a burden during this task.

The results are also promising, as our study task simulated a common scenario in real life—performing an action for hand-object interaction while focusing on another object or action somewhere else. Real-world examples include grabbing a water bottle or shifting gears while driving, or switching brushes while painting.

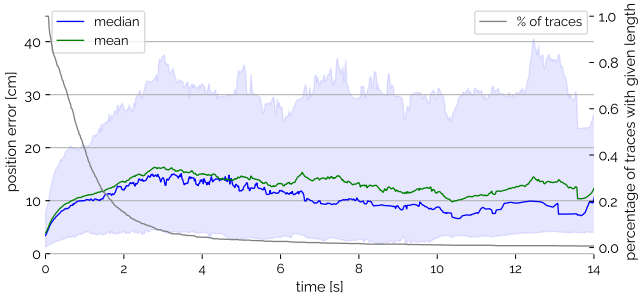

Figure 8: HOOV's tracking accuracy over time (compared to OptiTrack tracking with sub-mm accuracy) in Study 1. Participants' interactions outside the FOV mostly occurred within the first three seconds, during which HOOV achieved the best tracking performance and which covers the majority of interactions outside the user's field of view [63]. HOOV's tracking also remains stable afterwards, where the median and mean position error does not significantly exceed 15 cm. The mean (green) and median (blue) are computed across the set of traces corresponding to a given duration of time outside the FOV. The area shaded in blue indicates the error interval between the 5th and 95th percentile.

However, the gain in speed came at the cost of accuracy as expected. Under visual control, participants achieved higher success rates for grabbing and dropping targets than operating outside their FOV without vision, which has been quantified in previous experiments [63, 64, 70]. Note that our targets were world-anchored and static; while the experimenter ensured that participants did not move from the position on the floor, shifts in upper-body pose may already result in a considerable effect.

Upon closer inspection of where the outside FOV input events were least correct, we saw a disproportional amount of errors for the spheres and drop zones behind the user. In these cases, participants found it difficult to locate their hand outside their field of view correctly, possibly because these locations were outside participants' proprioceptive range and thus the area they would naturally interact with outside their FOV in other use cases. Related, in erroneous cases, HOOV failed to detect the correct object especially when participants placed their hand just between two targets. Then, even small deviations in estimated positions led to binary changes of target outcomes. We note that future systems may take this into account and potentially increase the separation between targets to correct for the tracking inaccuracies, introducing a dead band between targets where input events have no effect.

In terms of HOOV's potential to complement the tracking capabilities of headsets with limited FOVs, we first highlight its small median error of < 8° for all three simulated FOV conditions in Table 1. Several applications in VR benefit from stable orientation estimation, such as games (e.g., Beat Saber) or 3D interactive environments (e.g., for ray-casting). Next, we underline HOOV's capability to estimate *absolute positions* outside the headset's tracking FOV. Given a suitable interface design for virtual content outside the user's FOV, the median error of 7 cm—about the length of an index finger—offers sufficient tracking accuracy to navigate through a range of targets in the hemisphere defined by the arm's length. For this purpose, elements would need sufficient size and spacing that would allow users to traverse them in conjunction with the haptic feedback rendered by our wrist device.

While HOOV's tracking estimates are closest to the ground-truth within the first three seconds after the hand leaves the headset's FOV, HOOV showed promise in that its tracking error deteriorated only minimally under our simulated 90° tracking FOV condition. The mean position error of ~10 cm and the mean angle of the

Table 2: The table lists the median error (*median pos.*) and mean absolute error (*mean. pos*) of the tracked wrist position for HOOV and other approaches reported in the literature in cm. It also shows the sensors used for tracking the wrist position. Note that the reported values from the literature rely on a Kinect 2.0 as the ground-truth reference sensor. Compared to related approaches that evaluate their tracking performance on recorded data *offline*, we also analyze HOOV in an *interactive* study where participants complete tasks based on HOOV's tracking estimates.

Approach	Sensors	Evaluation	median pos.	mean pos.
ArmTrak (offline) [50]	9-axis IMU	offline	9.2	-
ArmTrak (real-time) [50]	9-axis IMU	offline	13.3	-
MUSE [49]	9-axis IMU	offline	8.8	-
MUSE (MR: grab-reach) [49]	9-axis IMU	offline	~15	-
LimbMotion [69]	9-axis IMU (+edge detector)	offline	8.9	-
RNN [60]	6-axis IMU	offline	15.4	16.7
HOOV—Study 1 (Ours)	6-axis IMU (+initial trajectory)	offline	**7.77**	**7.00**
HOOV—Study 2 (Ours)	6-axis IMU (+initial trajectory)	**interactive**	15.11	16.97

difference rotation of ~8° offer encouraging benefits for HOOV to work in conjunction with future headsets optimizing on power and form factor.

Regarding participants' qualitative feedback (Figure 9), their overall reception of performing the tasks in the OptiTrack condition was positive. 8 out of 12 participants reported that they preferred the OptiTrack condition over OCULUS, as their necks felt strained from repetitively turning around. Several participants also mentioned that this condition felt more efficient, even if we saw during the analysis that they tended to make more mistakes. In contrast, 4 of the 12 participants said that they would rather look at the object while performing the task, because it felt 'safer.'

Comparison to related approaches. HOOV is not directly comparable to other methods that estimate the wrist position from a single IMU, because we evaluated our system on a different dataset and HOOV receives an initial position of the hand as input. However, HOOV's relatively small median position error of 7 cm demonstrates the strength of our estimation pipeline to handle the accurate estimation of the wrist pose over short periods. Comparing the results to the values reported in the literature, HOOV performs 15.5% better than ArmTrak's offline algorithm [50], 11.7% better than the Particle Filter-based approach by Shen et al. [49], 12.7% better than LimbMotion [69], which uses an additional remote edge device, and 49.5% better than the RNN proposed by Wei et al. [60] (see Table 2).

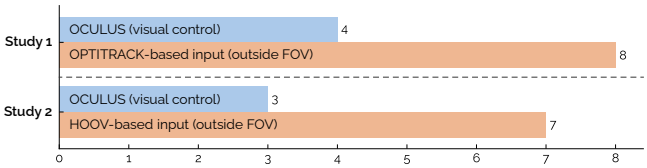

Figure 9: Participants' preferences of using one input technique over another. (Top) In Study 1, more participants preferred interacting in the OptiTrack condition than vice versa, as it did not require them to turn their heads. (Bottom) In Study 2, this trend remained, even though interaction outside the FOV was now driven by HOOV in real-time.

In terms of human performance, the results from our study generally align with the conclusions drawn by Yan et al. [64]. However, we also showed that participants were able to differentiate between objects in the lateral direction outside their field of view.

In order to understand HOOV's estimation performance during real-time use, we conducted another evaluation where participants' interaction was based solely on HOOV's tracking.

6 STUDY 2: EVALUATION WHERE PARTICIPANTS INTERACTED BASED ON HOOV'S REAL-TIME ESTIMATES

In Study 2, we conducted another evaluation where participants interacted based on HOOV's real-time predictions, such that any potential tracking error of our method affected their behavior and thus performance.

6.1 Study design

6.1.1 Apparatus. We built on the experiment apparatus from our previous study (Section 5.1). To additionally evaluate the accuracy of our pinch detection and position, we attached two motion capture markers to participants' thumb and index fingernails before the study. Tracking them, we simultaneously assessed the detection accuracy of pinch gestures as well as the accurate position of the fingers relative to the tracked wristband. Our apparatus detected pinch events when the two markers converged to a distance below 3 cm in a sudden motion.

In contrast to Study 1, the OptiTrack system was merely present in our apparatus to record ground-truth wrist poses for later analysis and had no influence on participants' behavior, the processed input events, or the virtual scene during the study. Likewise, the apparatus continued to detect pinch events based on the IMU signals (and not based on the attached OptiTrack markers).

The VR front-end of our study apparatus received the wrist position and orientation from HOOV's real-time estimations. For this, a PC (32 GB RAM, 3.60 GHz CPU) with an NVIDIA GeForce RTX 3090 received the headset's pose, hand positions from the VR headset when visible, and the stream of IMU signals from HOOV's wristband and processed them in real-time using the model trained on all data from Study 1 (subsection 5.2).

6.1.2 Participants. We recruited a fresh set of 10 participants (4 female, 6 male, ages 22–28, mean=25.3 years). Headset distances from the ground were 157—190 cm (mean=170 cm, SD=9 cm), and arm lengths ranged between 57—76 cm (mean=69 cm, SD=6 cm). Participants' prior experience with VR technology was lower than in our first study, ranging between 1–4 on a 5-point Likert scale (median=2). Participants received a small gratuity for their time.

6.1.3 Task. Participants completed the same three grab-and-place tasks (DROP, GRAB & COMPOUND) as in Study 1 while standing at an indicated stationary position. The placement of the 17 virtual spheres and drop zones was also identical to our first study.

6.1.4 Conditions. Participants again completed each task in two different conditions.

The first condition was identical to the OCULUS condition in Study 1 (Section 5.1), where participants visually followed their hands by turning their heads to interact with out-of-view targets.

In the second condition HOOV, participants faced straightforward and completed the tasks without turning their heads. Unlike in Study 1, our apparatus tracked participants' interaction, rendered positions, and executed input events based on HOOV's predictions. Participants again received haptic feedback at grab and release events and audio feedback depending on the outcome of a trial.

6.1.5 Procedure. At the start of the study, participants received an introduction to the tasks, provided information about their age and gender, and performed the T-pose calibration procedure to adjust the virtual environment to their body size. Participants received training for each condition and task that included visual feedback for the HOOV condition.

Overall, participants completed 272 trials (= 2 conditions × (34 DROP trials + 34 GRAB trials + 68 COMPOUND trials)) as fast as they could while avoiding errors. Each drop zone and sphere appeared as target twice for each combination of task and condition.

6.1.6 Measures. Using our study apparatus, we measured the duration of each trial as well as participants' success rates. We evaluated the tracking performance of HOOV compared to the ground-truth tracking from the OptiTrack's marker-based tracking system. Using the additional optical markers, we also analyzed the accuracy of our pinch gesture detection.

6.2 Results

6.2.1 Success rate. Using the OCULUS, participants correctly performed the task in 97.65% (max=100.00, min=91.18, SE=0.83) of all trials for DROP, 95.88% (max=100.00, min=85.29, SE=1.43) for GRAB, and 97.94% (max=100.00, min=92.65, SE=0.72) for COMPOUND.

Using HOOV's real-time tracking, participants correctly completed the task in 87.65% (max=97.06, min=70.59, SE=2.04) of trials for DROP, 86.18% (max=100.00, min=73.53, SE=2.28) for GRAB, and 86.18% (max=91.18, min=76.47, SE=1.23) for COMPOUND.

Analyzing the sub-mm accuracy measurements from the motion capture system for comparison, participants' hands were in the correct position 92.35% (max=97.06, min=76.47, SE=1.83) of the trials for DROP, 95.59% (max=100.00, min=79.41, SE=1.75) for GRAB and 92.21% (max=100.00, min=72.06, SE=2.09) during the COMPOUND task. Note that this analysis of human performance during interaction based on proprioception is hypothetical, as any real-time tracking error incurred by HOOV had an impact on participants' behavior during this study.

6.2.2 Trial duration. Using OCULUS, participants on average completed trials in 3.17 s (max=4.06, min=2.10, SE=0.20) for DROP, 3.74 s (max=5.26, min=2.48, SE=0.22) for GRAB, and 3.90 s (max=6.41, min=2.49, SE=0.32) for COMPOUND. Using HOOV's real-time tracking and without participants turning their heads, mean trial duration was 2.99 s (max=3.63, min=2.44, SE=0.11) for DROP, 3.54 s (max=4.07, min=2.93, SE=0.10) for GRAB, and 3.38 s (max=4.11, min=2.87, SE=0.13) for COMPOUND.

6.2.3 Pinch accuracy. Compared to the detected pinches using the two additional optical markers, HOOV's pinch detection achieved a recall of 94.85% and a precision of 98.05% across all three tasks. Any false-negative detection in HOOV during the study led participants to pinch again, since OptiTrack tracking had no impact on the behavior of the visual environment and study procedure.

6.2.4 Absolute tracking accuracy of HOOV. When considering HOOV as a continuous tracker of wrist position, HOOV's median position error was 15.11 cm (MAE=16.97 cm) averaged across all trials and participants. In terms of accuracy during input events outside the FOV, the position error of HOOV's predicted release and grab events was 14.43 (±4.16) cm for DROP, 16.41 (±4.21) cm for GRAB, and 14.85 (±3.19) cm for COMPOUND.

Analyzing HOOV's refined estimates of fingertip positions (15 cm offset to the wrist) using the markers attached to participants' fingernails, the average error was 4.9 cm when using the OptiTrack's wrist position as the basis and 14.0 cm when using HOOV's estimated wrist orientation and position as the basis for refinement (which encompasses HOOV's error of predicting wrist positions).

6.2.5 Qualitative feedback. As shown in Figure 9 (bottom), participants again preferred performing the study using HOOV overall, that is without requiring them to turn their heads to bring offscreen targets into view first. 7 out of the 10 participants reported a preference for the HOOV condition over OCULUS. Participants again mentioned the efficiency of not having to look for close-by offscreen targets before selecting them, although our analysis again showed that they had a slightly decreased success rate.

6.3 Discussion

Overall and compared to Study 1, participants were more careful in performing tasks in Study 2, both in the OCULUS and the HOOV condition. This is evident as the success rates for OCULUS and OptiTrack both exceeded participants' results in the first study using the OptiTrack system (see Figure 6). This also explains the longer average trial completion time in Study 2 as shown in Figure 7.

Specific to the HOOV condition, participants achieved a similar performance in terms of success rate. This demonstrates the capability of our learning-based method to generalize to and support interaction for users unseen during training without the need for calibration or fine-tuning. It also shows our method's suitability for proprioceptive interaction when possibly imperfect tracking influences user behavior and thus alters their interaction. Further supporting our method was participants' qualitative feedback, which highlighted that HOOV's tracking capabilities that extend and complement those of a regular headset were well-received.

In terms of tracking accuracy, HOOV produced a higher position error than in Study 1. Again, this likely resulted from the longer time that participants spent completing trials and, thus, the increased amount of time they interacted outside the field of view. This result is also commensurate with our assessment of tracking error over time as shown in Figure 8. Nevertheless, the absolute error in position remained relatively stable over time and still enabled participants to complete the intended tasks successfully.

The results from Study 2 also highlight the suitability of our pinch detection. Our system is tuned towards a more conservative prediction of pinch events, resulting in a low number of wrongly

detected peaks. This rarely required participants to pinch more than once when our online detection missed an event.

Finally, this evaluation also validated our previous assumption of a constant offset between HOOV's predicted wrist position and the participant's fingertip. The offset we found from the pinch position reported by the optical tracker was moderate and well below the error introduced by interacting based on proprioception outside the FOV in the first place.

Comparison to related approaches. The wrist position error for the interactive tracking in Study 2 is comparable to the results reported for Wei et al.'s RNN-based method [60]. However, their results were obtained through an offline evaluation (similar to Study 1) using a Kinect camera with limited accuracy, and the displayed user motions had a limited range in the yaw direction. While our performance falls short of the reported median tracking error of 8.8 cm for MUSE [49] in a static setting with unsupervised user motions, the error becomes comparable in a medical rehabilitation (MR) application involving reach-and-grasp tasks (see Table 2). In addition, MUSE relies on a 9-axis IMU with a magnetometer prone to interference and performs the tracking in a 5-DoF space with respect to the torso.

7 APPLICATIONS

Being complementary to the optical tracking abilities of commercial headsets, a multitude of applications can benefit from HOOV during operation to support outside FOV interaction. A large number of game titles and immersive experiences rely on external tracking to obtain the position of the user's hand-held controllers around the body. Such external tracking also supports comfortable interaction, where users can let their arms hang loosely by their sides, but still engage in an interactive experience thanks to controller input. With HOOV, we aim to enable similar levels of interactivity while supporting the portability-optimized form factor of current VR headsets, which therefore track all input inside-out as opposed to relying on additional sensing infrastructure. To that end, we prototyped three demonstrations that illustrate the benefits of HOOV in various application areas.

7.1 Beat Saber

To first illustrate the use of our system's extended tracking capabilities, we extended the *Beat Saber* gameplay to include omnidirectional targets. Rather than presenting targets to users from a single direction, our game scene contains four tracks. Since users do not have a full visual understanding of the arriving targets, they have to rely more greatly on cues from the music to execute movements. For instance, they may be encouraged to swing backward on a regular beat. With inside-out tracking capabilities of HMDs, these movements would be difficult to track, precluding such interactions from being implemented in current applications. HOOV resolves this issue and extends opportunities for gameplay to take advantage of a more dynamic range of movements.

7.2 Archery

Similar to Beat Saber, firing a bow requires movements that sometimes involve moving one's hand out of the tracking field-of-view of the head-mounted display. Unlike Beat Saber, archery further requires the user to steadily maintain an out-of-view hand position. HOOV nonetheless supports this and further enables the release interaction with the on-wrist IMU.

7.3 Block Builder

Lastly, we implemented a block builder application to demonstrate the applicability of HOOV in areas beyond games and play. The Block Builder application consists of a workbench that is instrumented with containers for blocks around the edges. Users can grab specific blocks from each container and piece them together on the workbench. As in a variety of other everyday activities, such as typing, we do not necessarily interact with the world with visual control, but rather sometimes rely on our sense of proprioception for control. As users gradually gain familiarity with the environment, we can expect users to rely more so on their spatial memory to grab items as opposed to visual control. HOOV enables this interaction.

8 LIMITATIONS AND FUTURE WORK

While the results of the method we introduced in this paper are promising, several limitations exist that deserve further investigation in future work.

HOOV as a method to compensate for IMU drift. Combating the drift arising from estimating positions based on observations from inertial motions is a challenging problem. With HOOV, we so far address this for the limited space of a person's reach, only for a short period of time, and only for stationary and standing scenarios. Figure 8 shows the development of tracking error over time, highlighting the challenge and remaining work needed to improve dead reckoning based on IMU signals. HOOV can therefore not be considered a general-purpose hand position estimator using inertial sensing. The results we presented in this paper also assume reaching and grabbing tasks and may therefore not translate to other tasks users may perform in their proprioceptive reach, such as gesturing, drawing, or other kinds of spatial navigation. The assumptions about the task, space, and scenarios we made to develop HOOV specifically for stationary VR scenarios considerably set apart the problem we addressed from more general inertial-based odometry, which attempts dead reckoning over much larger areas in mobile and moving settings.

Non-stationary environments. During our evaluation, participants stood still at a fixed position in the experimental space while performing all tasks as instructed. Though calibrated to each participant's body dimensions once at the beginning of the experiment, all targets were static and remained world-anchored. Although the experimenter verified throughout that participants did not step away from this position, the nature of the task did not allow us to control for a completely steady posture. Therefore, shifts in upper-body posture or slight shoulder rotations may have contributed to the errors we observed, as participants obtained no feedback about their position in the virtual space relative to the targets.

Future evaluations could extend our evaluation to include non-stationary environments where participants can move around. This would additionally evaluate the robustness of the method to motion artifacts. Because HOOV's algorithm currently needs a PC with a

Figure 10: Our demo applications demonstrate the benefits of our sensing method for three different use cases: (a) a user plays an adapted version of Beat Saber with objects arriving from four different directions that require a wide tracking range to support the intended movements, (b) while shooting an arrow users pull the bowstring far into the back requiring hand tracking outside the headset cameras' FOV, and (c) a block builder app where users can reach for objects outside their FOV.

good GPU, such an evaluation would require a mobile reimplementation of our method, using model-size reduction techniques and deployment on the embedded platform to support such operation.

Computational complexity. Due to the self-attention layers, HOOV's neural network incurs a complexity that grows quadratically with the length of the input sequence, limiting the maximum outside-the-field-of-view tracking duration. Future work should consider architectures with recurrent elements and memory [9] to avoid the repeated processing of the whole input sequence.

Tracking accuracy. While the tracking accuracy of HOOV is state-of-the-art for the given input modality, running from IMU-based orientation estimates only given the last hand observation from the headset, the existing error leaves room for improvement, especially compared to the tracking capabilities of an optical tracking system or a head-mounted display. We believe that a part of this challenge can be addressed by acquiring a much larger data set, which would allow our method to be refined for personalization and to adapt to varying body sizes.

Instrumentation of the wrist. HOOV requires the user to wear a wristband with an embedded IMU to be tracked outside the FOV and to submit input commands. While this is a common sensor inside every smartwatch today, it still requires a separate device to operate (parts of) VR experiences, which is counter to the intuition of manufacturers to integrate all VR components into just the headset. Given the existing ecosystem of watches and their various programming interfaces, we are excited about the possibility of transferring and adapting our method to run on commodity devices and to support out-of-the-box use. This will necessarily entail optimization strategies on the performance level as mentioned above in order to reduce the computational cost of our model or outsource its estimations to the neural computing unit of a headset.

9 CONCLUSION

We have presented HOOV, a wrist-worn sensing and position estimation method that complements existing AR and VR headsets in tracking the user's hand interactions outside the tracking field of view of their built-in cameras. HOOV obtains its input from a 6-axis IMU that is embedded in a wrist-worn band, which captures 3-axis acceleration and 3-axis rotational velocity. From these observations, HOOV's learning-based inference pipeline estimates the user's current wrist orientation and position as soon as the hand leaves the tracking range of the headset.

Our user studies showed the promise of our method, allowing participants to leverage their proprioception and interact with one out of 17 targets outside their FOV with an overall mean success rate of around 85% (Study 1) and 87% (Study 2). The outside FOV interaction in our study supported speed improvements over what would be required on today's standalone headsets—turning one's head to bring the virtual content into view first and exclusively interact with it *under visual control*. While under visual control, participants successfully interacted with the correct targets in 19 of 20 cases, we quantified the drop in success rates under outside FOV operation: Monitored with an external motion capture system, participants' success rate dropped to 18 of 20 trials when interacting purely based on their proprioception (and with targets that remained static and world-anchored). Of these 18 successful trials, ~17 interactions were correctly detected using HOOV—without any external infrastructure or cameras.

Our study also uncovered HOOV's potential as a tracking complement for raw 3D positions outside the headset's tracking FOV.

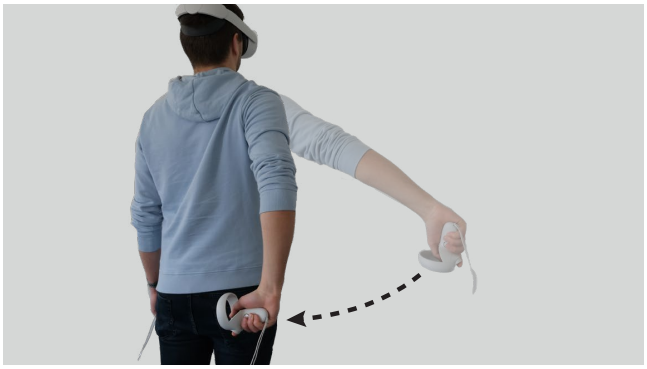

Figure 11: HOOV's wrist band contains the same sensor that is commonly found in hand-held controllers, allowing our method to generalize to a variety of input controllers.

HOOV's median tracking error of 7 cm for short-term outside-field-of-view interactions outperforms related techniques while requiring a commodity sensor that is commonly found in smartwatches and current hand-held controllers (Figure 11).

HOOV opens up an exciting opportunity to broaden the interactive space for AR and VR headsets that may now be able to better leverage users' proprioception within the large outside FOV space around the user that can accommodate convenient reach.

Collectively, we believe that HOOV will support the development of a future generation of AR and VR headsets that optimize for form factor and power consumption by outsourcing some of the computation to other devices. We see an opportunity for adaptive interaction techniques—as commonly used for selection tasks where input is ambiguous (e.g., touchscreen typing)—to operate in conjunction with HOOV's estimates to support users' proprioceptive interactions. Even more, we see the future potential for HOOV to support hand tracking and interaction *inside* the headset's FOV, thereby alleviating the computational cost of performing computer vision-based camera processing in real-time with a hand position estimation that may run on an embedded wrist-worn device.

ACKNOWLEDGMENTS

We sincerely thank Manuel Meier for helpful discussions and comments. We are grateful to NVIDIA for the provision of computing resources through the NVIDIA Academic Grant. We thank the anonymous reviewers and all participants of our user studies.

REFERENCES

[1] 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* (New Orleans, LA, USA) *(UIST '19)*. Association for Computing Machinery, New York, NY, USA, 453–462. https://doi.org/10.1145/3332165.3347889

[2] Karan Ahuja, Eyal Ofek, Mar Gonzalez-Franco, Christian Holz, and Andrew D Wilson. 2021. Coolmoves: User motion accentuation in virtual reality. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 2 (2021), 1–23.

[3] Karan Ahuja, Vivian Shen, Cathy Mengying Fang, Nathan Riopelle, Andy Kong, and Chris Harrison. 2022. ControllerPose: Inside-Out Body Capture with VR Controller Cameras. In *CHI Conference on Human Factors in Computing Systems*. 1–13.

[4] Jackie Andrade and Peter Meudell. 1993. Is spatial information encoded automatically in memory? *The Quarterly Journal of Experimental Psychology* 46, 2 (1993), 365–375.

[5] Alan Baddeley. 1992. Working Memory. *Science* 255, 5044 (1992), 556–559. http://www.jstor.org/stable/2876819

[6] Sina Bahram, Arpan Chakraborty, and Robert St. Amant. 2012. CAVIAR: A Vibrotactile Device for Accessible Reaching. In *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces* (Lisbon, Portugal) *(IUI '12)*. Association for Computing Machinery, New York, NY, USA, 245–248. https://doi.org/10.1145/2166966.2167009

[7] Gilles Bailly, Jörg Müller, Michael Rohs, Daniel Wigdor, and Sven Kratz. 2012. Shoesense: a new perspective on gestural interaction and wearable applications. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1239–1248.

[8] S James Biggs and Mandayam A Srinivasan. 2002. Haptic interfaces. *Handbook of virtual environments* (2002), 93–116.

[9] Aydar Bulatov, Yuri Kuratov, and Mikhail S Burtsev. 2022. Recurrent Memory Transformer. *arXiv preprint arXiv:2207.06881* (2022).

[10] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7291–7299.

[11] Ke-Yu Chen, Shwetak N Patel, and Sean Keller. 2016. Finexus: Tracking precise motions of multiple fingertips using magnetic sensing. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 1504–1514.

[12] Andy Cockburn and Bruce McKenzie. 2002. Evaluating the Effectiveness of Spatial Memory in 2D and 3D Physical and Virtual Environments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Minneapolis, Minnesota, USA) *(CHI '02)*. Association for Computing Machinery, New York, NY, USA, 203–210. https://doi.org/10.1145/503376.503413

[13] A. Cockburn, P. Quinn, C. Gutwin, G. Ramos, and J. Looser. 2011. Air pointing: Design and evaluation of spatial target acquisition with and without visual feedback. *International Journal of Human-Computer Studies* 69, 6 (2011), 401–414. https://doi.org/10.1016/j.ijhcs.2011.02.005

[14] David Dobbelstein, Philipp Hock, and Enrico Rukzio. 2015. Belt: An Unobtrusive Touch Input Device for Head-Worn Displays. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 2135–2138. https://doi.org/10.1145/2702123.2702450

[15] Jan Gugenheimer, David Dobbelstein, Christian Winkler, Gabriel Haas, and Enrico Rukzio. 2016. FaceTouch: Enabling Touch Interaction in Display Fixed UIs for Mobile Virtual Reality. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (Tokyo, Japan) *(UIST '16)*. Association for Computing Machinery, New York, NY, USA, 49–60. https://doi.org/10.1145/2984511.2984576

[16] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. 2018. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7297–7306.

[17] Sean Gustafson, Daniel Bierwirth, and Patrick Baudisch. 2010. Imaginary Interfaces: Spatial Interaction with Empty Hands and without Visual Feedback. In *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology* (New York, New York, USA) *(UIST '10)*. Association for Computing Machinery, New York, NY, USA, 3–12. https://doi.org/10.1145/1866029.1866033

[18] Sean Gustafson, Christian Holz, and Patrick Baudisch. 2011. Imaginary Phone: Learning Imaginary Interfaces by Transferring Spatial Memory from a Familiar Device. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology* (Santa Barbara, California, USA) *(UIST '11)*. Association for Computing Machinery, New York, NY, USA, 283–292. https://doi.org/10.1145/2047196.2047233

[19] Carl Gutwin, Andy Cockburn, and Nickolas Gough. 2017. A Field Experiment of Spatially-Stable Overviews for Document Navigation. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 5905–5916. https://doi.org/10.1145/3025453.3025905

[20] Jouni Hartikainen, Arno Solin, and Simo Särkkä. 2011. Optimal filtering with Kalman filters and smoothers. *Department of biomedica engineering and computational sciences, Aalto University School of Science, 16th August* (2011).

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*. 1026–1034.

[22] Shraga Hocherman. 1993. Proprioceptive guidance and motor planning of reaching movements to unseen targets. *Experimental brain research* 95, 2 (1993), 349–358.

[23] Yinghao Huang, Manuel Kaufmann, Emre Aksan, Michael J. Black, Otmar Hilliges, and Gerard Pons-Moll. 2018. Deep Inertial Poser Learning to Reconstruct Human Pose from SparseInertial Measurements in Real Time. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)* 37, 6 (Nov. 2018), 185:1–185:15.

[24] Du Q Huynh. 2009. Metrics for 3D rotations: Comparison and analysis. *Journal of Mathematical Imaging and Vision* 35 (2009), 155–164.

[25] Dong-Hyun Hwang, Kohei Aso, Ye Yuan, Kris Kitani, and Hideki Koike. 2020. Monoeye: Multimodal human motion capture system using a single ultra-wide fisheye camera. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 98–111.

[26] Northern Digital Inc. 2017. trakSTAR. https://www.ndigital.com/msci/products/drivebay-trakstar

[27] Hao Jiang and Kristen Grauman. 2017. Seeing invisible poses: Estimating 3d body pose from egocentric video. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 3501–3509.

[28] Jiaxi Jiang, Paul Streli, Huajian Qiu, Andreas Fender, Larissa Laich, Patrick Snape, and Christian Holz. 2022. Avatarposer: Articulated full-body pose tracking from sparse motion sensing. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*. Springer, 443–460.

[29] Haojian Jin, Christian Holz, and Kasper Hornbaek. 2015. Tracko: Ad-Hoc Mobile 3D Tracking Using Bluetooth Low Energy and Inaudible Signals for Cross-Device Interaction. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology (UIST '15)*. ACM, New York, NY, USA, 147–156. https://doi.org/10.1145/2807442.2807475

[30] Rudolf E. Kalman. 1960. A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering* 82, 1 (03 1960), 35–45. https://doi.org/10.1115/1.3662552 arXiv:https://asmedigitalcollection.asme.org/fluidsengineering/article-pdf/82/1/35/5518977/35_1.pdf

[31] Kibum Kim, Xiangshi Ren, Seungmoon Choi, and Hong Z. Tan. 2016. Assisting people with visual impairments in aiming at a target on a large wall-mounted display. *International Journal of Human-Computer Studies* 86 (2016), 109–120. https://doi.org/10.1016/j.ijhcs.2015.10.002

[32] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[33] Frank Chun Yat Li, David Dearman, and Khai N. Truong. 2009. Virtual Shelves: Interactions with Orientation Aware Devices. In *Proceedings of the 22nd Annual ACM Symposium on User Interface Software and Technology* (Victoria, BC, Canada) *(UIST '09)*. Association for Computing Machinery, New York, NY, USA, 125–128. https://doi.org/10.1145/1622176.1622200

[34] Frank Chun Yat Li, David Dearman, and Khai N. Truong. 2010. Leveraging Proprioception to Make Mobile Phones More Accessible to Users with Visual Impairments. In *Proceedings of the 12th International ACM SIGACCESS Conference on Computers and Accessibility* (Orlando, Florida, USA) *(ASSETS '10)*. Association for Computing Machinery, New York, NY, USA, 187–194. https://doi.org/10.1145/1878803.1878837

[35] Shuang Li, Jiaxi Jiang, Philipp Ruppel, Hongzhuo Liang, Xiaojian Ma, Norman Hendrich, Fuchun Sun, and Jianwei Zhang. 2020. A mobile robot hand-arm teleoperation system by vision and imu. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 10900–10906.

[36] Yang Liu, Chengdong Lin, Zhenjiang Li, Zhidan Liu, and Kaishun Wu. 2019. Poster: When Wearable Sensing Meets Arm Tracking. In *17th ACM International Conference on Mobile Systems, Applications, and Services, MobiSys 2019*. ACM New York, 518–519.

[37] WP Medendorp, S Van Asselt, and CCAM Gielen. 1999. Pointing to remembered visual targets after active one-step self-displacements within reaching space. *Experimental Brain Research* 125, 1 (1999), 50–60.

[38] Manuel Meier, Paul Streli, Andreas Fender, and Christian Holz. 2021. TapID: Rapid touch interaction in virtual reality using wearable sensing. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*. IEEE, 519–528.

[39] Damien Michel, Ammar Qammaz, and Antonis A Argyros. 2017. Markerless 3d human pose estimation and tracking based on rgbd cameras: an experimental evaluation. In *Proceedings of the 10th International Conference on PErvasive Technologies Related to Assistive Environments*. 115–122.

[40] Mark R. Mine, Frederick P. Brooks, and Carlo H. Sequin. 1997. Moving Objects in Space: Exploiting Proprioception in Virtual-Environment Interaction. In *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '97)*. ACM Press/Addison-Wesley Publishing Co., USA, 19–26. https://doi.org/10.1145/258734.258747

[41] Meta Motion. 2018. Gypsy Motion Capture System. http://metamotion.com/gypsy/gypsy-motion-capture-system.htm

[42] OptiTrack. 2022. Motion Capture Systems. http://optitrack.com/

[43] Ken Pfeuffer, Benedikt Mayer, Diako Mardanbegi, and Hans Gellersen. 2017. Gaze + Pinch Interaction in Virtual Reality. In *Proceedings of the 5th Symposium on Spatial User Interaction* (Brighton, United Kingdom) *(SUI '17)*. Association for Computing Machinery, New York, NY, USA, 99–108. https://doi.org/10.1145/3131277.3132180

[44] Albert Postma and Edward HF De Haan. 1996. What was where? Memory for object locations. *The Quarterly Journal of Experimental Psychology Section A* 49, 1 (1996), 178–199.

[45] Ivan Poupyrev and Tadao Ichikawa. 1999. Manipulating Objects in Virtual Worlds: Categorization and Empirical Evaluation of Interaction Techniques. *Journal of Visual Languages & Computing* 10, 1 (1999), 19–35. https://doi.org/10.1006/jvlc.1998.0112

[46] Helge Rhodin, Christian Richardt, Dan Casas, Eldar Insafutdinov, Mohammad Shafiei, Hans-Peter Seidel, Bernt Schiele, and Christian Theobalt. 2016. EgoCap: Egocentric Marker-Less Motion Capture with Two Fisheye Cameras. *ACM Trans. Graph.* 35, 6, Article 162 (nov 2016), 11 pages. https://doi.org/10.1145/2980179.2980235

[47] Angelo Maria Sabatini. 2011. Kalman-filter-based orientation determination using inertial/magnetic sensors: Observability analysis and performance evaluation. *Sensors* 11, 10 (2011), 9182–9206.

[48] Joey Scarr, Andy Cockburn, and Carl Gutwin. 2013. Supporting and Exploiting Spatial Memory in User Interfaces. *Found. Trends Hum.-Comput. Interact.* 6, 1 (dec 2013), 1–84. https://doi.org/10.1561/1100000046

[49] Sheng Shen, Mahanth Gowda, and Romit Roy Choudhury. 2018. Closing the gaps in inertial motion tracking. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*. 429–444.

[50] Sheng Shen, He Wang, and Romit Roy Choudhury. 2016. I am a smartwatch and i can track my user's arm. In *Proceedings of the 14th annual international conference on Mobile systems, applications, and services*. 85–96.

[51] Charles S Sherrington. 1907. On the proprio-ceptive system, especially in its reflex aspect. *Brain* 29, 4 (1907), 467–482.

[52] John F Soechting and Martha Flanders. 1989. Sensorimotor representations for pointing to targets in three-dimensional space. *Journal of neurophysiology* 62, 2 (1989), 582–594.

[53] Hans Strasburger. 2020. Seven Myths on Crowding and Peripheral Vision. *i-Perception* 11, 3 (2020). https://doi.org/10.1177/2041669520913052 PMID: 32489576.

[54] Paul Streli, Jiaxi Jiang, Andreas Rene Fender, Manuel Meier, Hugo Romat, and Christian Holz. 2022. TapType: Ten-finger text entry on everyday surfaces via Bayesian inference. In *CHI Conference on Human Factors in Computing Systems*. 1–16.

[55] Md. Sami Uddin, Carl Gutwin, and Benjamin Lafreniere. 2016. HandMark Menus: Rapid Command Selection and Large Command Sets on Multi-Touch Displays. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5836–5848. https://doi.org/10.1145/2858036.2858211

[56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[57] Daniel Vlasic, Rolf Adelsberger, Giovanni Vannucci, John Barnwell, Markus Gross, Wojciech Matusik, and Jovan Popović. 2007. Practical motion capture in everyday surroundings. *ACM transactions on graphics (TOG)* 26, 3 (2007), 35–es.

[58] Dong-Bach Vo and Stephen A. Brewster. 2015. Touching the invisible: Localizing ultrasonic haptic cues. In *2015 IEEE World Haptics Conference (WHC)*. 368–373. https://doi.org/10.1109/WHC.2015.7177740

[59] Timo Von Marcard, Bodo Rosenhahn, Michael J Black, and Gerard Pons-Moll. 2017. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. In *Computer Graphics Forum*, Vol. 36, No. 2. Wiley Online Library, 349–360.

[60] Wenchuan Wei, Keiko Kurita, Jilong Kuang, and Alex Gao. 2021. Real-time 3D arm motion tracking using the 6-axis IMU sensor of a smartwatch. In *2021 IEEE 17th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*. IEEE, 1–4.

[61] Alexander Winkler, Jungdam Won, and Yuting Ye. 2022. QuestSim: Human Motion Tracking from Sparse Sensors with Simulated Avatars. In *SIGGRAPH Asia 2022 Conference Papers*. 1–8.

[62] Erwin Wu, Ye Yuan, Hui-Shyong Yeo, Aaron Quigley, Hideki Koike, and Kris M. Kitani. 2020. Back-Hand-Pose: 3D Hand Pose Estimation for a Wrist-Worn Camera via Dorsum Deformation Network. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '20)*. Association for Computing Machinery, New York, NY, USA, 1147–1160. https://doi.org/10.1145/3379337.3415897

[63] Huiyue Wu, Yanyi Deng, Jiajun Pan, Tianxing Han, Yonglin Hu, Kaini Huang, and Xiaolong Luke Zhang. 2021. User capabilities in eyes-free spatial target acquisition in immersive virtual reality environments. *Applied Ergonomics* 94 (2021), 103400.

[64] Yukang Yan, Chun Yu, Xiaojuan Ma, Shuai Huang, Hasan Iqbal, and Yuanchun Shi. 2018. Eyes-Free Target Acquisition in Interaction Space around the Body for Virtual Reality. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173616

[65] Xinyu Yi, Yuxiao Zhou, Marc Habermann, Soshi Shimada, Vladislav Golyanik, Christian Theobalt, and Feng Xu. 2022. Physical Inertial Poser (PIP): Physics-aware Real-time Human Motion Tracking from Sparse Inertial Sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13167–13178.

[66] Xinyu Yi, Yuxiao Zhou, and Feng Xu. 2021. TransPose: real-time 3D human translation and pose estimation with six inertial sensors. *ACM Transactions on Graphics (TOG)* 40, 4 (2021), 1–13.

[67] Yang Zhang, Chouchang Yang, Scott E Hudson, Chris Harrison, and Alanson Sample. 2018. Wall++ room-scale interactive and context-aware sensing. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–15.

[68] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. 2018. Through-wall human pose estimation using radio signals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7356–7365.

[69] Han Zhou, Yi Gao, Xinyi Song, Wenxin Liu, and Wei Dong. 2019. Limbmotion: Decimeter-level limb tracking for wearable-based human-computer interaction. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 4 (2019), 1–24.

[70] Qiushi Zhou, Difeng Yu, Martin N Reinoso, Joshua Newn, Jorge Goncalves, and Eduardo Velloso. 2020. Eyes-free target acquisition during walking in immersive mixed reality. *IEEE Transactions on Visualization and Computer Graphics* 26, 12 (2020), 3423–3433.

[71] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. 2019. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5745–5753.