2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Advancing Virtual Reality Interaction: A Ring-Shaped Controller and Pose Tracking

Zhuqing Zhang[1,2*], Dongxuan Li[2], Jiayao Ma[2], Yijia He[2], Pan Ji[2], Rong Xiong[1], Hongdong Li[3], Yue Wang[1]

***Abstract*— Ensuring robust tracking of controllers' movement is critical for human-robot interaction in virtual reality (VR) scenarios. This paper proposes a robust tracking algorithm based on a novel wearable ring-shaped controller equipped with an inertial measurement unit (IMU) and a light-emitting diode (LED). This novel controller design allows users to free up their hands for more immersive experiences. To track the controller's motion accurately and robustly, we resort to various forms of visual measurements, including 6 DoF and 5 DoF pose measurements from hand gesture detection, as well as 3 DoF position measurement and 2 DoF image measurement derived from the LED. We theoretically analyze the performances of these observation models and propose an optimal observation model combination scheme. Moreover, the necessity and rationale of online estimating system gravity are illustrated. The effectiveness of our tracking method is validated through extensive experiments.**

I. Introduction

An increasing amount of attention has been focused on achieving more natural and intuitive human-robot interaction, particularly in performing human-robot collaborative tasks. To protect people from exposure to dangerous and harsh envrionments, a practical solution is to interact with the robot remotely via a virtual or augmented reality (VR/AR) setting [1]–[3]. As with VR games, hand controllers are typically used to facilitate human-robot interaction. By tracking the movement of the controllers in 3D space, the user's action can be mapped to the world, enabling the user to interact with the environment. However, how to define the form of the controller and track the controller robustly remains an open question. This paper introduces a novel easy-to-use controller, as well as a robust tracking method (using both vision and inertial observations) for accurately estimating the pose of the controller.

Exisiting methods for tracking the 6 DoF motions of controllers can be classified into outside-in methods and inside-out methods. Traditional outside-in solutions rely on pre-installed and fixed base-stations [6]–[8], limiting usage scenarios. In contrast, inside-out tracking schemes usually rely on built-in cameras in the controllers [9], [10] to actively locate their position via simultaneous localization and mapping (SLAM) [11], [12]. While such inside-out tracking schemes offers better flexibility, the additional camera sensors can pose challenges to the product's power consumption and data transmission bandwidth.

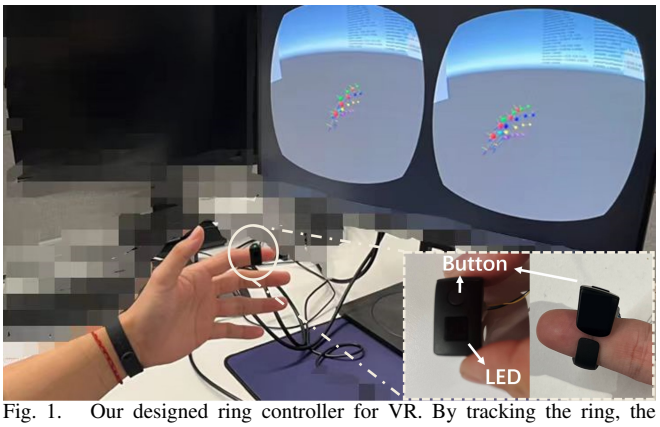


Fig. 1. Our designed ring controller for VR. By tracking the ring, the reconstructed hand can be rendered in the virtual environment.

A solution that lies between outside-in and inside-out tracking schemes is increasingly popular in VR products [4], [5]. This type of tracking scheme usually utilizes cameras in the head-mounted display (HMD) to track the controller by detecting the LEDs embedded in the controller. However, most existing VR controllers are handheld, which may negatively impact the user experience by limiting certain actions like grabbing and pinching. In contrast, directly detecting and tracking the bare hand seems to be a plausible solution to circumvent these drawbacks. However, while many works have conducted detailed studies on hand pose estimation [20], [21], producing accurate and robust estimations remains a challenging task. Therefore, it is preferable to design a wearable controller that allows users to free their hands for a more immersive experience, while also being able to track hands accurately and robustly with sensors in the controller.

The usually employed wearable devices include gloves [13], [14] and smart armband [15]. But they tend to be equipped with multiple sensors, such as multiple IMUs [16], bend or flex sensors [14], [17], and external cameras [13], making the interaction cost too high to reach the consumer level (refer to recent surveys [18], [19] for more details).

In this paper, we design a wearable ring controller that only requires one 6-axis IMU and one LED (cf. Fig. 1). The controller is not only comfortable to wear but also has low cost and power consumption. Based on the de-

[1]State Key Laboratory of Industrial Control Technology and Institute of Cyber-Systems and Control, Zhejiang University, Zhejiang, China.
[2]Tencent XR Vision Labs, Tencent Holdings, Shenzhen, China.
[3]College of Engineering and Computer Science, Australian National University, Canberra.
*This work was done while interning at Tencent XR Vision Labs.
This work was partly supported by the National Nature Science Foundation of China under Grant 62373322.
Yijia He and Yue Wang are the co-corresponding authors (heyijia2016@gmail.com, wangyue@iipc.zju.edu.cn).

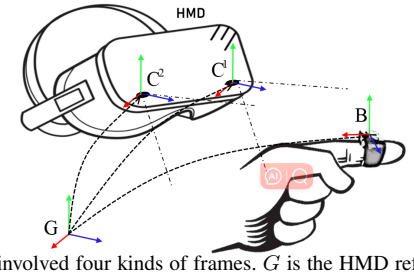


Fig. 2. The involved four kinds of frames. G is the HMD reference frame; C^i is the HMD camera frame; B is the IMU frame in the ring.

signed controller, we propose a robust tracking algorithm that utilizes various observation models, including 2 DoF image observation and 3 DoF position observation derived from the LED, and 5 DoF and 6 DoF pose observations derived from a learning-based hand pose detection method [23]. The advantages and disadvantages of these observation models are thoroughly analyzed in this paper. In particular, the potential degeneration scenario of the 2 DoF observation is discussed in detail. Drawing on these analyses, we propose an optimal observation combination scheme. Furthermore, a method for online estimation of the system's gravity is proposed to improve the tracking performance and the rationale behind this method is analyzed theoretically.

In summary, the contributions of this paper include:

- Propose a lightweight and low-cost ring-shaped controller for VR interaction scenes, and design a robust tracking algorithm for the system, including various observation models from 2 DoF to 6 DoF.
- A comparative study of different observation models is performed in detail. Particularly, the potential degeneration issue of the LED observation model is theoretically analyzed and solved. According to the comparative study, an optimal observation model combination scheme is given.
- The rationale of online estimating the gravity vector is demonstrated by theoretic analysis, and the corresponding gravity estimation method is proposed.

II. System Overview

A. Designed Controller

Figure 1 illustrates our designed ring-shaped controller. The ring has a diameter of 2cm and features an opening at the bottom to accommodate various finger sizes. It incorporates a button for facilitating user interaction with the virtual environment. To track the controller, only a 6-axis IMU and an LED are deployed. In comparison to commonly used handheld controllers, our ring-shaped controller is compact, lightweight, and low-cost, offering users a better immersion.

B. Problem Statement

Apart from the ring-shaped controller, our tracking system also includes an HMD (cf. Fig. 2). The HMD locates its position by the built-in IMU and cameras via a SLAM algorithm. The rest of the paper focuses on how to track the 6 DoF pose of the ring utilizing the embedded IMU and LED of the controller and the cameras of HMD.

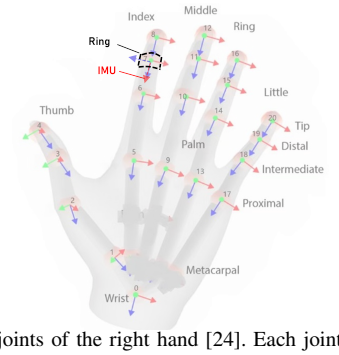


Fig. 3. The 21 joints of the right hand [24]. Each joint has a number ID.

As indicated in Fig. 2, there are three types of coordinate systems. Frame G represents the HMD inertial reference frame. The gravity vector in G is ${}^G\mathbf{g} = [0, 0, 9.8]^\top$. C^i $(i = 1, 2)$ is the camera frame of HMD. B is the IMU (body) frame of the controller. Our aim is to estimate the pose of the controller in G.

C. State Space Modeling

The basic system state at time step k is defined as[1]:

$$\mathbf{x}_k = \begin{bmatrix} {}^G\mathbf{q}_{B_k}^\top & {}^G\mathbf{p}_{B_k}^\top & {}^G\mathbf{v}_{B_k}^\top & \mathbf{b}_{a_k}^\top & \mathbf{b}_{g_k}^\top \end{bmatrix}^\top, \quad (1)$$

where ${}^G\mathbf{q}_{B_k}$ is a unit quaternion, representing the orientation of B in G. Its corresponding rotation matrix ${}^G\mathbf{R}_{B_k} \in SO(3)$ transforms a 3-dimensional vector from B to G. ${}^G\mathbf{p}_{B_k} \in \mathbb{R}^3$ is the position of B in G. ${}^G\mathbf{v}_{B_k} \in \mathbb{R}^3$ is the velocity of B in G. $\mathbf{b}_{a_k}$ and $\mathbf{b}_{g_k}$ are the biases of accelerometer and gyroscope in IMU, respectively.

With the state defined in (1), the error state is given by:

$$\tilde{\mathbf{x}}_k = \begin{bmatrix} {}^G\tilde{\boldsymbol{\theta}}_{B_k}^\top & {}^G\tilde{\mathbf{p}}_{B_k}^\top & {}^G\tilde{\mathbf{v}}_{B_k}^\top & \tilde{\mathbf{b}}_{a_k}^\top & \tilde{\mathbf{b}}_{g_k}^\top \end{bmatrix}^\top, \quad (2)$$

where ${}^G\tilde{\boldsymbol{\theta}}_{B_k} = \mathrm{Log}({}^G\hat{\mathbf{R}}_{B_k}^\top \, {}^G\mathbf{R}_{B_k})$. $\mathrm{Log}(\cdot)$ is the logarithmic operation on $SO(3)$, mapping $SO(3)$ to $\mathbb{R}^3$. The other elements are defined by the standard error in $\mathbb{R}^3$.

D. State Estimation

Given computational constraints, we apply the lightweight error-state Kalman filter (ESKF) [22] for state estimation. To initialize the system, we need to give the initial pose of the ring ${}^G\mathbf{T}_{B_0}$. ${}^G\mathbf{T}_{B_0}$ can be acquired through ${}^G\hat{\mathbf{T}}_B = {}^G\hat{\mathbf{T}}_{C^1} {}^{C^1}\hat{\mathbf{T}}_B$, where ${}^G\hat{\mathbf{T}}_{C^1}$ is from the HMD SLAM algorithm, and ${}^{C^1}\hat{\mathbf{T}}_B$ is from the hand detection algorithm [23]. At this point, we can propagate the state in G with IMU measurements and update the state with the different observation models following the ESKF procedure.

III. Observation Models of Hand and Ring

A. Observation Models based on Hand Gesture Detection

6 DoF Observation Model: When the HMD observes the user's hand, the hand detection algorithm can produce the 3D positions of the 21 joints (cf. Fig. 3) in frame C^1. To generate the pose of the ring, we first use the pair of joints with

[1]In this paper, the variable with $\hat{\cdot}$, $\tilde{\cdot}$, and $\bar{\cdot}$ represent the estimation, the error, and the observation, respectively. The variable without superscript represents the ground truth.

IDs $\{6, 14\}$ and $\{6, 8\}$ to perform Schmidt orthogonalization [36], where we assume the direction vector derived from the joints 6 and 8 is the x axis of the ring, such that the orientation of the ring is obtained. The position of the joint 7 is regarded as the position of the ring. Then, we can obtain the observation of the ring pose in frame G, ${}^{G}\bar{\mathbf{T}}_{B_k} = {}^{G}\hat{\mathbf{T}}_{C_k^1}{}^{C^1}\bar{\mathbf{T}}_{B_k}$, leading to the observation function:

$$ {}^{G}\bar{\mathbf{R}}_{B_k} = {}^{G}\mathbf{R}_{B_k}\mathrm{Exp}(\mathbf{n}_{R_k}),\ {}^{G}\bar{\mathbf{p}}_{B_k} = {}^{G}\mathbf{p}_{B_k} + \mathbf{n}_{p_k}, \quad (3) $$

where $\mathbf{n}_R$ and $\mathbf{n}_p$ denote the observation Gaussian white noise for the orientation and position parts, respectively. $\mathrm{Exp}(\cdot)$ is an exponential function on $SO(3)$ that maps $\mathbb{R}^3$ to $SO(3)$. Then, the observation error function can be computed: $\mathbf{r}_k^{6D} = \mathbf{H}_k^{6D}\tilde{\mathbf{x}}_{k|k-1} + \mathbf{n}_k^{6D}$, where $\mathbf{r}$, $\mathbf{H}$, and $\mathbf{n}$ are observation error, Jacobian matrix of the observation function and observation noise, respectively. The right superscript $6D$ represents that the corresponding variables are based on the 6 DoF model.

5 DoF Observation Model: In (3), the orientation observations of the ring ${}^{G}\bar{\mathbf{R}}_B$ are derived from joint positions with IDs $6, 8$, and 14. However, since joint 14 is likely to be blocked, and the two fingers are not necessarily coplanar, the obtained orientation is susceptible to detection noise, resulting in inaccurate orientation estimation. In contrast, the direction vector derived from joints 6 and 8 is more appropriate for representing the ring orientation. Based on this fact, we introduce the following function to impose a 2 DoF constraint on the ring orientation:

$$ 1 = {}^{G}\bar{\mathbf{p}}_{6,8_k}^{\top}({}^{G}\mathbf{R}_{B_k}\mathbf{l}_x) + n_k \quad (4) $$

where ${}^{G}\bar{\mathbf{p}}_{6,8_k}$ is a unit vector computed from the positions of joints 6 and 8 in the frame G. $\mathbf{l}_x = [1, 0, 0]^{\top}$ so that ${}^{G}\mathbf{R}_{B_k}\mathbf{l}_x$ is the x-axis of the ring frame B in G. n is the direction observation noise. In (4), we assume that ${}^{G}\bar{\mathbf{p}}_{6,8_k}$ and ${}^{G}\mathbf{R}_{B_k}\mathbf{l}_x$ should be parallel (as indicated by Fig. 3).

By combining (4) with the position observation, we can get the observation error: $\mathbf{r}_k^{5D} = \mathbf{H}_k^{5D}\tilde{\mathbf{x}}_{k|k-1} + \mathbf{n}_k^{5D}$.

B. Observation Models based on the LED

3 DoF Observation Model: When the two cameras in the HMD observe the LED, the 3D position of the LED in C^1 can be triangulated. By combining the HMD pose, we can obtain the LED position in G: ${}^{G}\bar{\mathbf{p}}_{L_k} = {}^{G}\hat{\mathbf{R}}_{C_k^1}{}^{C^1}\hat{\mathbf{p}}_{L_k} + {}^{G}\hat{\mathbf{p}}_{C_k^1}$. Because the ring is small enough, the extrinsic between the LED and the IMU can be regarded as an identity matrix. Therefore, the observation function can be expressed as:

$$ {}^{G}\bar{\mathbf{p}}_{L_k} = {}^{G}\mathbf{p}_{B_k} + \mathbf{n}_{p_{L_k}}, \quad (5) $$

where $\mathbf{n}_{p_L}$ is the observation noise. Using (5), we can get the 3 DoF observation error: $\mathbf{r}_k^{3D} = \mathbf{H}_k^{3D}\tilde{\mathbf{x}}_{k|k-1} + \mathbf{n}_k^{3D}$.

2 DoF Observation Model: Apart from triangulating the LED 3D position, we can reproject the LED 3D position into the HMD images, leading to the 2D observation model:

$$ \mathbf{z}_k^i = \Pi^i({}^{C^i}\mathbf{p}_{L_k}) + \mathbf{n}_{z_k} = \Pi^i({}^{G}\mathbf{R}_{C_k^i}^{\top}({}^{G}\mathbf{p}_{B_k} - {}^{G}\mathbf{p}_{C_k^i})) + \mathbf{n}_{z_k}, \quad (6) $$

where $\mathbf{z}^i$ is the pixel observation in the image captured by the HMD camera i. $\Pi^i(\cdot)$ is the projection function of the camera i. $\mathbf{n}_{z_k}$ is the observation noise. Based on (6), we can get the 2 DoF observation error: $\mathbf{r}_k^{2D} = \mathbf{H}_k^{2D}\tilde{\mathbf{x}}_{k|k-1} + \mathbf{n}_k^{2D}$. The dimension of $\mathbf{r}^{2D}$ can be two or four, depending on the number of cameras that can observe the LED.

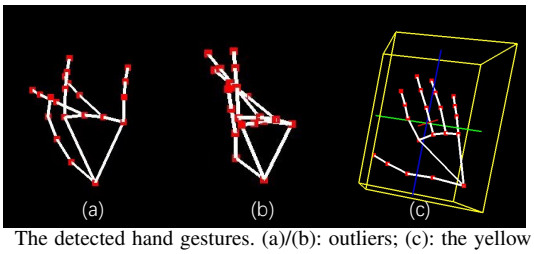


Fig. 4. The detected hand gestures. (a)/(b): outliers; (c): the yellow box is derived from PCA.

IV. Comparative Study on Different Models

A. Pros and Cons of Different Models

5 DoF vs. 6 DoF: As indicated in Sec. III-A, the derived 3D orientation of the ring is susceptible to the noise. On the contrary, the parallel assumption indicated by (4) is more plausible. Therefore, 5 DoF observation model is preferred, which will be validated in Sec. VI.

3 DoF vs. 2 DoF: In comparison to the 2 DoF observation model, the 3 DoF observation model requires that the LED is within the field of view of both cameras simultaneously. On the contrary, the 2 DoF model does not have this limitation.

However, compared with the 3 DoF model, the 2 DoF model exhibits strong nonlinearity because of the introduction of the projection function $\Pi^i(\cdot)$. Under the framework of ESKF, if the linearization point (estimated state) is far away from the ground truth, the linear approximation to the strong nonlinear model can be poor, leading to the divergence of the system. As shown in Sec. VI, this problem is likely to happen in VR: When the LED is out of sight for a while and reappears, only relying on the IMU measurements will result in poor estimation. Under this situation, when the LED measurements become available again, it is preferable to utilize the 3 DoF observation model or employ the method in [25] to change a more accurate linearization point.

5/6 DoF vs. 2/3 DoF: Usually, it is more favorable to employ the $5/6$ DoF model than the $2/3$ DoF model as the former provides higher dimensional constraints, and hands have better visibility than the LED. However, as the $5/6$ DoF observations are derived from the learning-based method [23], outliers may occur when the hand gesture is unfriendly (cf. Fig. 4 (a) and (b)). In this case, we should opt for the $2/3$ DoF observation model. To identify outliers of hand detection, we design the following rules: 1) From the perspective of human ergonomics, the finger composed of the detected joints should form a convex shape (anti-joint is unacceptable). Besides, the angle between two adjacent fingers should be less than a certain threshold; 2) Some special gestures (e.g., fist) will also lead to poor results because of finger occlusion. To eliminate such outliers, we employ Principal Component Analysis (PCA) (cf. Fig. 4 (c)). By analyzing the ratio between the two primary directions of the estimated hand pose and comparing it with a given threshold, the outliers are filtered.

Degeneration Cases of 2 DoF: To investigate the degeneration case of the 2 DoF observation model, it is necessary to analyze the observability of the system [26]. For simplicity, we neglect the IMU bias of $\mathbf{x}_k$ in (1), resulting in a simplified state as $\mathbf{x}'_k = \begin{bmatrix} {}^G\mathbf{q}_{B_k}^\top & {}^G\mathbf{p}_{B_k}^\top & {}^G\mathbf{v}_{B_k}^\top \end{bmatrix}^\top$. With the state propagation function and the 2 DoF observation model, the observability matrix of the system from the time step m to k can be given by:

$$\mathcal{M} \triangleq \begin{bmatrix} \mathbf{H}_m^{2D} \\ \mathbf{H}_{m+1}^{2D}\mathbf{\Phi}_{m+1|m} \\ \vdots \\ \mathbf{H}_k^{2D}\mathbf{\Phi}_{k|m} \end{bmatrix} = \begin{bmatrix} \Gamma_m^i[\mathbf{0}_{3\times3} & {}^G\mathbf{R}_{C_m^i}^\top & \mathbf{0}_{3\times3}] \\ & \vdots & \\ \Gamma_k^i[\mathbf{M}_k & {}^G\mathbf{R}_{C_k^i}^\top & {}^G\mathbf{R}_{C_k^i}^\top \Delta t_{k|m}] \end{bmatrix} \tag{7}$$

where $\mathbf{\Phi}_{k|m}$ is the state transition matrix from time step m to k, $\mathbf{M}_k = -{}^G\mathbf{R}_{C_k^i}\lfloor {}^G\mathbf{p}_{B_k} - {}^G\mathbf{p}_{B_m} - {}^G\mathbf{v}_{B_m}\Delta t_{k|m} + \frac{1}{2}{}^G\mathbf{g}\Delta t_{k|m}^2 \rfloor_\times {}^G\mathbf{R}_{B_m}$, $\lfloor\cdot\rfloor_\times$ transforms a 3D vector to a skew symmetry matrix, $\Delta t_{k|m}$ is the time gap between m and k, Γ_k^i is the Jacobian matrix of the projection function $\Pi^i(\cdot)$ with respect to ${}^{C^i}\mathbf{p}_{L_k}$ (cf. (6)). Denote ${}^{C^i}\mathbf{p}_{L_k} \triangleq [x_k^i, y_k^i, z_k^i]^\top$ and suppose the focus parameters of the camera i are f_x^i, f_y^i, then $\Gamma_k^i = \begin{bmatrix} \frac{f_x^i}{z_k^i} & 0 & -\frac{f_x^i x_k^i}{z_k^{i2}} \\ 0 & \frac{f_y^i}{z_k^i} & -\frac{f_y^i y_k^i}{z_k^{i2}} \end{bmatrix}$.

Under the normal situation, $\mathcal{M}$ is column-full-rank, indicating that the state of the system from m to k is *fully observable*. However, when the ring remains static, the observability of the system will degenerate.

• a) *When the ring is static:* For the ideal case, ${}^G\mathbf{v}_{B_m} = \cdots = {}^G\mathbf{v}_{B_k} = \mathbf{0}, {}^G\mathbf{p}_{B_m} = \cdots = {}^G\mathbf{p}_{B_k} \triangleq \mathbf{p}, {}^G\mathbf{R}_{B_m} = \cdots = {}^G\mathbf{R}_{B_k} \triangleq \mathbf{R}$, then $\mathbf{M}_k$ degenerates to $-{}^G\mathbf{R}_{C_k^i}\lfloor \frac{1}{2}{}^G\mathbf{g}\Delta t_{k|m}^2 \rfloor_\times \mathbf{R}$. We can find that, in this case, $\mathcal{M}$ has the following right null space:

$$\mathcal{N}_1 = \begin{bmatrix} ({}^W\mathbf{g})^\top & (\mathbf{0}_{3\times1})^\top & (\mathbf{0}_{3\times1})^\top \end{bmatrix}^\top, \tag{8}$$

which implies that *when the ring is stationary, the direction along the gravity becomes unobservable*. The same problem can also occur when using the 3 DoF observation model. The derivation is omitted here since it is quite similar to the case of the 2 DoF observation model.

• b) *When the ring and the HMD are static:* In this case, the observation of the ring under the HMD camera i keeps unchanged, such that ${}^{C^i}\mathbf{p}_{L_m} = \cdots = {}^{C^i}\mathbf{p}_{L_k} \triangleq \mathbf{p}_L = [x^i, y^i, z^i]^\top$, ${}^G\mathbf{R}_{C_m^i} = \cdots = {}^W\mathbf{R}_{C_k^i} \triangleq \mathbf{R}_{C^i}$. Therefore, $\Gamma_m^i = \cdots = \Gamma_k^i \triangleq \Gamma^i$. Denoting $\mathbf{n} = [x^i/z^i, y^i/z^i, 1]^\top$, apart from $\mathcal{N}_1$, $\mathcal{M}$ has another one-dimensional right null space:

$$\mathcal{N}_2 = \begin{bmatrix} \mathbf{0}_{3\times1}^\top & (\mathbf{R}_C^i\mathbf{n})^\top & \mathbf{0}_{3\times1}^\top \end{bmatrix}^\top, \tag{9}$$

which indicates that *when both the ring and the HMD are static, one direction of the position becomes unobservable*. It is worth noting that in (7), only one camera observation is considered, which is the root cause of the generation of $\mathcal{N}_2$. If two cameras observe the LED simultaneously, this degeneration case will vanish.

However, we argue that degeneration case b) can happen in practice, whose impact is more significant than that of degeneration case a). This is because, in the state propagation stage, the position part is the second-order integral of the IMU measurements (i.e., linear acceleration), whereas the orientation part is the first-order integral of the IMU measurements (i.e., angular velocity). This fact makes the position more sensitive to the IMU measurement noises. If the observation fails to constrain the acceleration/velocity/position of the system, the position part would drift rapidly.

To solve the degeneration problem, an intuitive and effective method is zero velocity update (ZUPT) [31].

B. Combination of Observation Models

Based on the analyses mentioned above, we make a reasonable combination of the four kinds of observation models to make the tracking system more accurate and robust. The logic of this combination is as follows:

- The 6 DoF observation model is not used because the 5 DoF model could provide more accurate constraints.
- The 5 DoF observation is employed as the primary source to update the system state because the user's hands have much better visibility than the LED.
- When the 5 DoF measurements are considered outliers, the LED-based measurements are employed to update the system state if they are available.
- When the LED-based measurements are employed, we prefer to employ the 3 DoF observation model due to its better linearity. If the LED is observed only by a single camera of the HMD, the 2 DOF observation model will be used instead.
- In the rare event that both the 5 DoF measurements are outliers and the LED is blocked, we choose to utilize the outliers to update the state rather than discard them. Although outliers usually have poor orientation, the position of the joint 7 can be satisfied. which can prevent the system from significant drift caused by relying only on IMU measurements.

The observation combination logic can be summarized as: high quality 5 DoF→3 DoF→2 DoF→low quality 5 DoF.

V. Online Gravity Estimation

To accurately estimate the state in frame G, it is imperative to acquire a precise gravity vector ${}^G\mathbf{g}$. In the above introduction, we assume G to be a perfect inertial frame so that ${}^G\mathbf{g} = [0, 0, 9.8]^\top$. However, in practice, G is likely to be an imperfect inertial frame due to the estimation errors in the intialization procedure. In such cases, assuming ${}^G\mathbf{g} = [0, 0, 9.8]^\top$ would compromise the accuracy of state estimation. An intuitive solution to this problem is to optimize ${}^G\mathbf{g}$ online. In this section, we will analyze the rationale behind this approach in our system.

A. State Space Modeling

Suppose the perfect inertial frame is denoted as $\bar{G}$ and the actual imperfect inertial frame as G. Then, ${}^G\hat{\mathbf{g}} = {}^G\hat{\mathbf{R}}_{\bar{G}}{}^{\bar{G}}\mathbf{g}$ holds, where ${}^{\bar{G}}\mathbf{g} = [0, 0, 9.8]^\top$. Hence, optimizing ${}^G\mathbf{g}$ is equivalent to optimizing ${}^G\hat{\mathbf{R}}_{\bar{G}}$. Since the yaw angle is unobservable for an inertial frame, we only need to optimize

the pitch and roll angles of ${}^{G}\hat{\mathbf{R}}_{\bar{G}}$. To be specific, we extend the system state and system error defined in (1) and (2) as:

$$\mathbf{x}_k^* = \begin{bmatrix} {}^{G}\mathbf{q}_{B_k}^\top & {}^{G}\mathbf{p}_{B_k}^\top & {}^{G}\mathbf{v}_{B_k}^\top & \mathbf{b}_{a_k}^\top & \mathbf{b}_{g_k}^\top & {}^{G}\mathbf{q}_{\bar{G}_k}^\top \end{bmatrix}^\top, \quad (10)$$

$$\tilde{\mathbf{x}}_k^* = \begin{bmatrix} {}^{G}\tilde{\boldsymbol{\theta}}_{B_k}^\top & {}^{G}\tilde{\mathbf{p}}_{B_k}^\top & {}^{G}\tilde{\mathbf{v}}_{B_k}^\top & \tilde{\mathbf{b}}_{a_k}^\top & \tilde{\mathbf{b}}_{g_k}^\top & {}^{G}\tilde{\boldsymbol{\theta}}_{\bar{G}_k}^\top \end{bmatrix}^\top. \quad (11)$$

It is worth noting that ${}^{G}\tilde{\boldsymbol{\theta}}_{\bar{G}_k}$ represents the error of *roll and pitch angles* of ${}^{G}\mathbf{q}_{\bar{G}_k}$, which is a two-dimensional vector.

The kinematic differential equation of the system is then given as:

$$\begin{cases} {}^{G}\dot{\mathbf{q}}_{B_k} = {}^{G}\mathbf{q}_{B_k} \otimes [0, \frac{1}{2}(\bar{\boldsymbol{\omega}}_k - \mathbf{b}_{g_k} - \mathbf{n}_{g_k})^\top]^\top & {}^{G}\dot{\mathbf{p}}_{B_k} = {}^{G}\mathbf{v}_{B_k} \\ {}^{G}\dot{\mathbf{v}}_{B_k} = {}^{G}\mathbf{R}_{B_k}(\bar{\mathbf{a}}_k - \mathbf{b}_{a_k} - \mathbf{n}_{a_k}) - {}^{G}\mathbf{R}_{\bar{G}_k}\,{}^{\bar{G}}\mathbf{g} & \\ \dot{\mathbf{b}}_{a_k} = \mathbf{n}_{ba_k} \quad \dot{\mathbf{b}}_{g_k} = \mathbf{n}_{bg_k} & {}^{G}\dot{\mathbf{q}}_{\bar{G}_k} = \mathbf{0} \end{cases}, \quad (12)$$

where $\otimes$ is the quaternion-based multiplication, $\mathbf{n}_{ba}$ and $\mathbf{n}_{bg}$ are the gaussian random walk noise.

B. Observability Analysis

While there are some inertial systems online optimizing the gravity estimation [27], [28], the rationale has not been theoretically analyzed. In fact, for an inertial SLAM system, there are ideally four unobservable directions [29]. Online optimization of gravity may reduce the unobservable directions to three, resulting in inconsistent estimations. In this section, we aim to investigate the rationale of online estimating gravity for our tracking system.

As mentioned in Sec. IV-A, our original tracking system is *fully observable*. If online estimating the gravity does not introduce unobservable subspaces into the system, then it is rational to perform online gravity estimation.

Consider the simplified state $\mathbf{x}_k^{*\prime} = \begin{bmatrix} {}^{G}\mathbf{q}_{B_k}^\top & {}^{G}\mathbf{p}_{B_k}^\top & {}^{G}\mathbf{v}_{B_k}^\top & {}^{G}\mathbf{q}_{\bar{G}_k}^\top \end{bmatrix}^\top$. Using (12) and (3), we can derive the following observation matrix[2]:

$$\mathcal{M}^{*\prime} = \begin{bmatrix} \mathbf{H}_m^{6D} \\ \mathbf{H}_{m+1}^{6D}\boldsymbol{\Phi}_{m+1|m}^{*\prime} \\ \vdots \\ \mathbf{H}_k^{6D}\boldsymbol{\Phi}_{k|m}^{*\prime} \end{bmatrix} = \begin{bmatrix} \mathbf{0}_{3\times3} & \mathbf{I}_{3\times3} & \mathbf{0}_{3\times3} & \mathbf{0}_{3\times2} \\ \mathbf{I}_{3\times3} & \mathbf{0}_{3\times3} & \mathbf{0}_{3\times3} & \mathbf{0}_{3\times2} \\ & & \vdots & \\ \mathbf{M}_1^{*\prime} & \mathbf{I}_{3\times3} & \mathbf{I}_{3\times3}\Delta t_{k|m} & \mathbf{M}_2^{*\prime} \\ \mathbf{M}_3^{*\prime} & \mathbf{0}_{3\times3} & \mathbf{0}_{3\times3} & \mathbf{0}_{3\times2} \end{bmatrix},$$

$$\mathbf{M}_1^{*\prime} = -\lfloor {}^{G}\mathbf{p}_{B_k} - {}^{G}\mathbf{p}_{B_m} - {}^{G}\mathbf{v}_{B_m}\Delta t_{k|m} + \frac{1}{2}{}^{G}\mathbf{R}_{\bar{G}_m}\,{}^{\bar{G}}\mathbf{g}\Delta t_{k|m}^2 \rfloor_\times {}^{G}\mathbf{R}_{B_m},$$

$$\mathbf{M}_2^{*\prime} = \frac{1}{2}({}^{G}\mathbf{R}_{\bar{G}_m}\lfloor {}^{\bar{G}}\mathbf{g} \rfloor_\times)_{first2cols}\Delta t_{k|m}^2, \quad \mathbf{M}_3^{*\prime} = {}^{B_k}\mathbf{R}_{B_m}. \quad (13)$$

Given that the system is *fully observable* when not optimizing the gravity online, we know that the first nine columns of $\mathcal{M}^{*\prime}$ are linearly independent. Furthermore, the last two columns of $\mathcal{M}^{*\prime}$ are both column-full-rank and linearly independent of the first nine columns. Therefore, we can conclude that online estimation of gravity does not introduce any unobservable subspaces into the system. As a result, the system remains *fully observable*, indicating the reasonability of online estimating the gravity.

[2]All the Jacobians are computed with the ground truth. Even though the 6 DoF observation model is employed here, we can extend the conclusion to the other observation models.

TABLE I

THE RMSE/CM OF THE ESTIMATED RING POSITION W./W.O. ONLINE GRAVITY OPTIMIZATION

Seq.	Online gravity estimation	5 DoF	6 DoF	RMSE
HO-1		✓		1.866
	✓	✓		1.610
			✓	1.806
	✓		✓	1.634
HO-2		✓		2.704
	✓	✓		2.315
			✓	2.826
	✓		✓	2.313
HL-E		✓		8.036
	✓	✓		1.990
			✓	7.015
	✓		✓	2.532
HL-M		✓		5.208
	✓	✓		2.562
			✓	6.155
	✓		✓	4.257
HL-H		✓		7.217
	✓	✓		5.325
			✓	7.192
	✓		✓	6.062

✓: the corresponding configuration is selected.

VI. EXPERIMENTS

Because the ring-shaped controller is rarely recorded in literature, and different controllers have different measurement sources[3], it is hard to compare the performance of our controller with other existing controllers. In this section, we conduct thorough ablation experiments to verify the conclusions obtained from the above analyses. We collected five data sequences as a user interacted with VR scenarios:

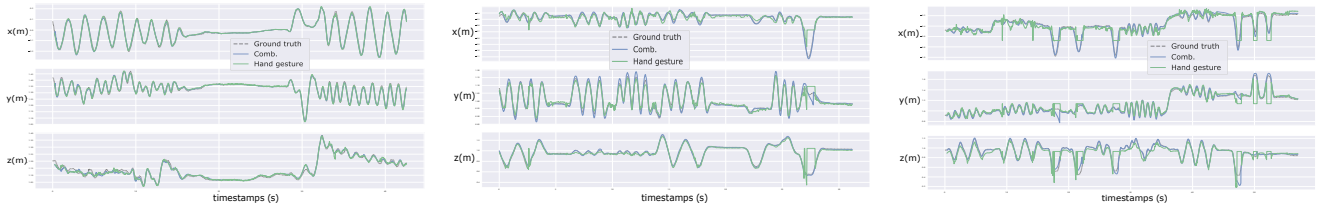
- Sequences HO-1 and HO-2 (**H**and pose **O**nly): These two sequences were recorded without short-exposure images, which made it impossible to detect the LED. As a result, only 5 DoF and 6 DoF measurements are available for these sequences.
- Sequences HL-E, HL-M, HL-H (**H**and pose and **L**ED): These three sequences have access to all four types of measurements. The sequence HL-E is considered to be an **E**asy one, as the hand can be continuously detected (cf. Fig. 6-a) and the gestures are well-behaved, with minimal chances of hand pose outliers. Sequence HL-M is considered to be of **M**oderate difficulty, as the hand will occasionally be out of sight, and there are more hand pose outliers. Sequence HL-H is deemed to be the **H**ardest of the three, as there are more peroids that the hand and the LED are out of sight, leading to the lowest measurement quality (cf. Fig. 6-c).

The ground truth data for the datasets were obtained from a motion capture device[4]. The timestamp alignment was performed following the way in [30].

Gravity Estimation: According to the analysis in Sec. V, online optimization of gravity is essential and has the theoretical foundation. To demonstrate the impact of online

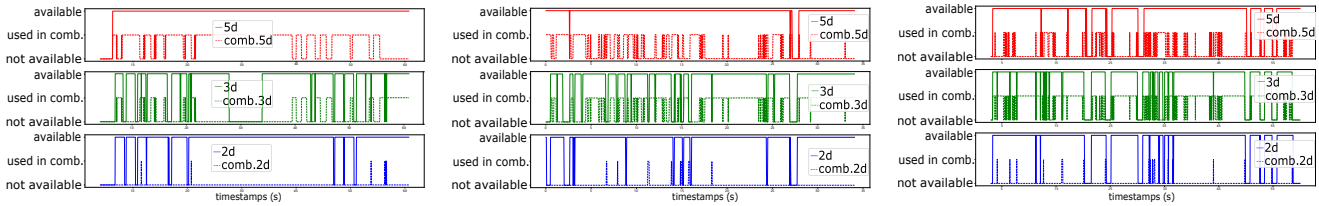
[3]To the best of our knowledge, existing research primarily concentrated on gesture recognition using rings [32], [33]. The ring proposed in [34] is equipped with different sensors from us. Additionally, the commercially available ring controllers [35] only offer 3 degrees of freedom (3DoF) information, with undisclosed details regarding their localization approach.

[4]Because the ring is too small, it can only attach one reflective ball, so the motion capture can only provide the ground truth of the position. However, we argue that the estimated position is coupled with the estimated orientation, i.e., good position accuracy indicates good orientation accuracy.



(a) Seq. HL-E (b) Seq. HL-M (c) Seq. HL-H

Fig. 5. Trajectory comparison. The dotted line is the ground truth; The blue one is the trajectory derived from our proposed combined scheme; The green one is from the hand gesture estimation algorithm [23].



(a) Seq. HL-E (b) Seq. HL-M (c) Seq. HL-H

Fig. 6. The availablility of different observation types (the solid line) vs. the utilization of different observation models under the combination (comb.) scheme (the dotted line) on different sequences.

optimizing gravity on system estimation accuracy, we conducted ablation studies on five captured sequences using 5 DoF and 6 DoF observation models. The root mean squared errors (RMSE) of the estimated ring positions, measured in centimeters (cm), are presented in Table I. The results show that online optimization of gravity can significantly improve the tracking system's accuracy under different models.

Observation Models Comparison: To validate the conclusions drawn in Section IV, we conducted experiments to evaluate the performance of the tracking system using different observation models, including our proposed combination scheme, while optimizing gravity online. The results of the experiments are presented in Table II.

TABLE II

THE RMSE/CM OF THE ESTIMATED RING POSITION WITH DIFFERENT OBSERVATION MODELS

Seq.	2 DoF	3 DoF	5 DoF	6DoF	Comb.
HO-1	—	—	**1.610**	1.634	—
HO-2	—	—	2.315	**2.313**	—
HL-E	2.893	2.485	1.998	2.532	**1.837**
HL-M	4.461	3.425	2.562	4.257	**2.290**
HL-H	5.040	**4.309**	5.345	6.062	4.901

—: there is no such measurement.

From Table II, the 5 DoF model produces better estimation than the 6 DoF model in almost all sequences. Therefore, we argue that the 5 DoF observation model is preferable for practical use.

When it comes to LED-based measurements, the performance of the 3 DoF model is better than the 2 DoF model. This is because the 3 DoF model is linear as analysized previously. When the LED is temporarily out of sight and observed again, the 3 DoF model will have much better convergence than the 2 DoF model. Therefore, the 3 DoF observation model is preferable to the 2 DoF model.

When we combine the different observation models by our proposed combination scheme, the system performance is improved. Particularly, compared to only relying on the 5 DoF observations, the results of Comb. are better, indicating the effectiveness of the proposed 5 DoF outlier rejection scheme. In Fig. 5, we compare the results from hand gesture estimation [23] (the green line) and our combined approach. The plots indicate that only relying on hand gesture detection, the trajectories are unsmooth. Especially when the user's hand is out of sight (cf. the flat areas of the green line). On the contrary, thanks to the IMU, hand pose, and LED measurements, the combined approach can produce smoother and more accurate results. The availability of different observation sources in sequences HL is plotted in Fig. 6 with solid lines. The observation models used in the combination scheme at each time step are plotted with dotted lines, revealing that 5 DoF, 3 DoF, and 2 DoF observations all contribute to the final results.

Degeneration Validation: For HL-E, there is a brief period when the controller and the HMD are stationary, and only a single camera of the HMD observes the LED. In this case, the special degeneration of the 2 DoF observation model analyzed in Sec. IV-A is triggered. For the result of HL-E—2 DoF in Table II, the ZUPT is employed during this period; otherwise, the result would be $6.569\ cm$. Fig. 7 gives the trajectory comparison along the z direction to show the effect of the degeneration. This result is consistent with our theoretical analysis.

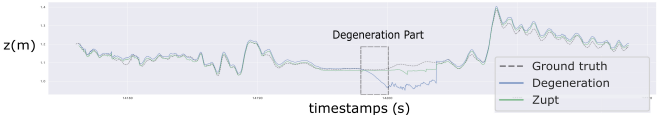


Fig. 7. Trajectory comparsion for demonstrating the degeneration case. Only 2 DoF observation is used to update the state.

VII. CONCLUSION

This paper introduces a lightweight and low-cost wearable controller for human-computer interaction in VR scenarios. The four types of observation models are investigated in detail based on this controller. According to theoretical analyses and experiments, an optimal scheme is developed to track the controller accurately and robustly. Future work includes investigating and solving corner cases to further improve the tracking system.

REFERENCES

[1] Q Wang, et al., "Modeling of human welders' operations in virtual reality human–robot interaction." IEEE Robotics and automation letters 4.3 (2019): 2958-2964.

[2] R Etzi, et al., "Using virtual reality to test human-robot interaction during a collaborative task." International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. Vol. 59179. American Society of Mechanical Engineers, 2019.

[3] M. Dianatfar, L. Jyrki, and L. Minna, "Review on existing VR/AR solutions in human–robot collaboration." Procedia CIRP 97 (2021): 407-411.

[4] Meta: Quest. https://www.meta.com/quest, 2023.

[5] Pimax: Crystal. https://pimax.com/crystal/, 2023.

[6] J. Lwowski, A. Majumdat, P. Benavidez, J. J. Prevost and M. Jamshidi, "HTC Vive Tracker: Accuracy for Indoor Localization," in IEEE Systems, Man, and Cybernetics Magazine, vol. 6, no. 4, pp. 15-22, Oct. 2020, doi: 10.1109/MSMC.2020.2969031.

[7] L.G. Sansone, R. Stanzani, M. Job, et al. "Robustness and static-positional accuracy of the SteamVR 1.0 virtual reality tracking system," Virtual Reality 26, 903–924, 2022. https://doi.org/10.1007/s10055-021-00584-5.

[8] S. Scheggi, L. Meli, C. Pacchierotti and D. Prattichizzo, "Touch the virtual reality: using the leap motion controller for hand tracking and wearable tactile devices for immersive haptic rendering," In ACM SIGGRAPH 2015 Posters (pp. 1-1). https://doi.org/10.1145/2787626.2792651.

[9] X. Jiang, et al. "A SLAM-based 6DoF controller with smooth auto-calibration for virtual reality." The Visual Computer, 1-14, 2022.

[10] T. Babic, R. Harald and H. Michael, "Pocket6: A 6dof controller based on a simple smartphone application," Proceedings of the 2018 ACM Symposium on Spatial User Interaction, pp. 2-10, 2018.

[11] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. M. Montiel and J. D. Tardós, "ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual–Inertial, and Multimap SLAM," in IEEE Transactions on Robotics, vol. 37, no. 6, pp. 1874-1890, Dec. 2021, doi: 10.1109/TRO.2021.3075644.

[12] P. Geneva, K. Eckenhoff, W. Lee, Y. Yang and G. Huang, "OpenVINS: A Research Platform for Visual-Inertial Estimation," *2020 IEEE International Conference on Robotics and Automation (ICRA)*, Paris, France, 2020, pp. 4666-4672, doi: 10.1109/ICRA40945.2020.9196524.

[13] A. T. Maereg, et al, "A low-cost, wearable Opto-Inertial 6-DOF hand pose tracking system for VR," Technologies, 5.3, 49, 2017.

[14] O. Glauser, et al, "Interactive hand pose estimation using a stretch-sensing soft glove," ACM Transactions on Graphics (ToG), 38.4, 1-15, 2019.

[15] De Paolis, T. Lucio and V. De Luca, "The impact of the input interface in a virtual environment: the Vive controller and the Myo armband," Virtual Reality 24.3, 483-502, 2020.

[16] Noitom: Hi5 Glove, https://hi5vrglove.com, 2023.

[17] Y. Zheng, et al, "Development and evaluation of a sensor glove for hand function assessment and preliminary attempts at assessing hand coordination," Measurement, 93, 1-12, 2016.

[18] W. Chen, et al, "A Survey on Hand Pose Estimation with Wearable Sensors and Computer-Vision-Based Methods," Sensors, 20(4):1074, https://doi.org/10.3390/s20041074.

[19] H. Kim, et al, "Recent Advances in Wearable Sensors and Integrated Functional Devices for Virtual and Augmented Reality Applications," Adv. Funct. Mater. 2021, 31, 2005692. https://doi.org/10.1002/adfm.202005692.

[20] A. Ahmad, M. Cyrille and D. Albert, "Hand pose estimation and tracking in real and virtual interaction: A review," Image and Vision Computing 89 (2019): 35-49.

[21] R. Li, Z. Liu and J. Tan, "A survey on 3D hand pose estimation: Cameras, methods, and datasets," Pattern Recognition 93 (2019): 251-272.

[22] Sola, Joan, "Quaternion kinematics for the error-state Kalman filter." arXiv preprint arXiv:1711.02508 (2017).

[23] S. Han, et al, "MEgATrack: monochrome egocentric articulated hand-tracking for virtual reality," ACM Trans. Graph. 39, 4, Article 87 (August 2020), 13 pages. https://doi.org/10.1145/3386569.3392452

[24] PICO Unity Integration SDK. https://developer-cn.pico-interactive.com/document/unity/hand-tracking, 2024

[25] Z. Zhang, Y. Jiao, S. Huang, R. Xiong and Y. Wang, "Map-Based Visual-Inertial Localization: Consistency and Complexity," in IEEE Robotics and Automation Letters, vol. 8, no. 3, pp. 1407-1414, March 2023, doi: 10.1109/LRA.2023.3239314.

[26] Ding, Xiaqing, et al. "Degeneration-aware localization with arbitrary global-local sensor fusion." Sensors 21.12 (2021): 4042.

[27] S. Lynen, M. W. Achtelik, S. Weiss, M. Chli and R. Siegwart, "A robust and modular multi-sensor fusion approach applied to MAV navigation," 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2013, pp. 3923-3929, doi: 10.1109/IROS.2013.6696917.

[28] W. Xu and F. Zhang, "FAST-LIO: A Fast, Robust LiDAR-Inertial Odometry Package by Tightly-Coupled Iterated Kalman Filter," in IEEE Robotics and Automation Letters, vol. 6, no. 2, pp. 3317-3324, April 2021, doi: 10.1109/LRA.2021.3064227.

[29] S. Jia, Y. Jiao, Z. Zhang, R. Xiong and Y. Wang, "FEJ-VIRO: A Consistent First-Estimate Jacobian Visual-Inertial-Ranging Odometry," 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Kyoto, Japan, 2022, pp. 1336-1343, doi: 10.1109/IROS47612.2022.9981413.

[30] D. Schubert, T. Goll, N. Demmel, V. Usenko, J. Stückler and D. Cremers, "The TUM VI Benchmark for Evaluating Visual-Inertial Odometry," 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 2018, pp. 1680-1687, doi: 10.1109/IROS.2018.8593419.

[31] Y. Wang, A. Chernyshoff and A. M. Shkel, "Study on Estimation Errors in ZUPT-Aided Pedestrian Inertial Navigation Due to IMU Noises," in IEEE Transactions on Aerospace and Electronic Systems, vol. 56, no. 3, pp. 2280-2291, June 2020, doi: 10.1109/TAES.2019.2946506.

[32] C. Liang, et al. "DualRing: Enabling subtle and expressive hand interaction with dual IMU rings." Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 5.3 (2021): 1-27.

[33] H. Zhou, et al. "Learning on the Rings: Self-Supervised 3D Finger Motion Tracking Using Wearable Sensors." Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 6.2 (2022): 1-31.

[34] H. Zhou, et al. "One Ring to Rule Them All: An Open Source Smartring Platform for Finger Motion Analytics and Healthcare Applications." Proceedings of the 8th ACM/IEEE Conference on Internet of Things Design and Implementation. 2023.

[35] Nolo. https://www.nolovr.com/News/Info/752, 2023.

[36] Gram-Schmidt Process. https://en.wikipedia.org/wiki/Gram-Schmidt_process, 2024.