SENSORS

# Visual-inertial hand motion tracking with robustness against occlusion, interference, and contact

**Yongseok Lee[1], Wonkyung Do[2], Hanbyeol Yoon[1], Jinuk Heo[1], WonHa Lee[3], Dongjun Lee[1]***



**State-of-the-art technologies for hand (and finger) motion tracking do not always provide accurate and robust tracking. For example, severe occlusions can affect tracking with vision sensors, electromagnetic interference affects tracking with inertial measurement units (IMUs) and compasses, and ambiguous mechanical contact can affect tracking with soft sensors (i.e., the inability to distinguish motion-induced deformation). Here, we report a visual-inertial skeleton tracking (VIST) framework that provides robust and accurate hand tracking in a variety of real-world scenarios. Our proposed VIST framework comprises a sensor glove with multiple IMUs and passive visual markers as well as a head-mounted stereo camera. VIST also uses a tightly coupled filtering-based visual-inertial fusion algorithm to estimate the hand/finger motion and autocalibrates hand/glove-related kinematic parameters simultaneously while taking into account the hand anatomical constraints. Our VIST framework exhibits good tracking accuracy and robustness, affordable material cost, lightweight hardware and software, and durability to permit washing. We validate our VIST framework through quantitative and qualitative experiments in real-world conditions. Our approach to hand tracking has the potential to enrich not only human-robot interaction applications (e.g., direct humanoid hand teleoperation, hand-based collaborative robot programming, and drone swarm control) but also the user experience in many virtual reality and augmented reality applications.**

## INTRODUCTION

Dexterous use of hands (with fingers) is one defining characteristic of humans. Replicating dexterity of the human hand would markedly improve efficiency, intuitiveness, and richness of many real-world human-robot interaction (HRI) applications, including (i) robotic hand teleoperation (Fig. 1B), particularly that of anthropomorphic robotic hands (*1*, *2*), where a remote user can fully use their hand and fingers with haptic feedback for complex manipulation tasks, instead of relying on [typically only up to 6 DOFs (degrees of freedom)] conventional haptic devices (*3*); (ii) collaborative robot interaction (Fig. 1C and Movie 1), where a user can quickly and intuitively provide rich commands and cues to the robot using their hands and fingers, thereby making the interaction safer and smoother compared to the case of conventional pendant programming (*4*); and (iii) 3D (three-dimensional) drone swarm control (Fig. 1D and Movie 2), where a user in the field can efficiently control the complex 3D swarm behavior by simply nudging their formation or quickly defining 3D virtual walls to avoid dangerous regions. All the tasks mentioned above are difficult when relying on conventional 2D tablet interfaces (*5*). This use of the hand will also greatly improve the user experience of virtual reality (VR) and augmented reality (AR), which is currently dominated by 6-DOF "fist-based" controllers (*6*, *7*).

Hand (with fingers) tracking is a key technology to enable the hand to be used in HRI, AR, or VR applications. We detail three approaches, with their respective fundamental limitations, which have been proposed to solve the hand tracking problem:

1) Vision-based hand tracking (*8*–*15*) uses cameras [e.g., red-green-blue (RGB), RGB-D (depth), or stereo] to track hand motion without markers while exploiting machine-learning techniques trained with large image datasets (*16*–*18*). However, the fundamental issue of occlusions [or outside camera field of view (FOV)] cannot be circumvented even with the machine-learning techniques, which are well known for issues of data dependency and generalization problems against hands, objects, and lighting conditions outside the training sets (*19*).

2) Inertial measurement unit (IMU)/compass-based wearable hand tracking (*20*–*24*) typically uses six-axis IMUs (i.e., accelerometer and gyroscope) and compasses (i.e., magnetometer). These are attached to each bone of the hand to measure its 3-DOF (absolute) orientation, and the hand configuration is reconstructed by collecting this angle information of each bone with additional hand position sensors [e.g., (*25*)]. However, this method is fundamentally susceptible to the magnetic field change or interference and thus impossible to use when near or in contact with ferromagnetic objects or electronic devices.

3) Soft wearable hand tracking (*26*–*30*) uses a number of soft sensors, each producing signals according to their deformations. These are wrapped around the hand to estimate the hand configurations with additional hand pose sensors [e.g., (*25*)]. However, this approach fundamentally suffers from the inability to distinguish motion-induced deformation from those induced by contact, rendering it unsuitable for applications where the user needs to handle objects/tools or wear haptic devices (*22*). Furthermore, there are difficulties involving calibration [due to sensor hysteresis and intersensor coupling (*31*)] and limited ruggedness [e.g., large bending or difficulty washing (*26*)].

Other approaches include pure magnetic trackers (*32*–*35*), which are susceptible to the eletromagnetic interference just as for the case of IMU/compass-based tracking, and exoskeletons (*36*–*39*), which require bulky [e.g., weights of 300 to 500 g (*36*–*38*)] and rigid mechanical structures, substantially compromising hand agility (e.g., friction) and long-term wearability (e.g., fatigue).

Here, we propose a visual-inertial skeleton tracking (VIST) system and its algorithm for accurate, robust, and affordable hand tracking

[1]Department of Mechanical Engineering, Seoul National University, 1 Gwanak-Ro, Gwanak-Gu, Seoul 08826, Republic of Korea. [2]Department of Mechanical Engineering, Stanford University, 438 Panama Street, Building 570, Stanford, CA 94305, USA. [3]Memory Business, Samsung Electronics, 114 Samsung-Ro, Gyeonggi-Do 17786, Republic of Korea.
*Corresponding author. Email: djlee@snu.ac.kr

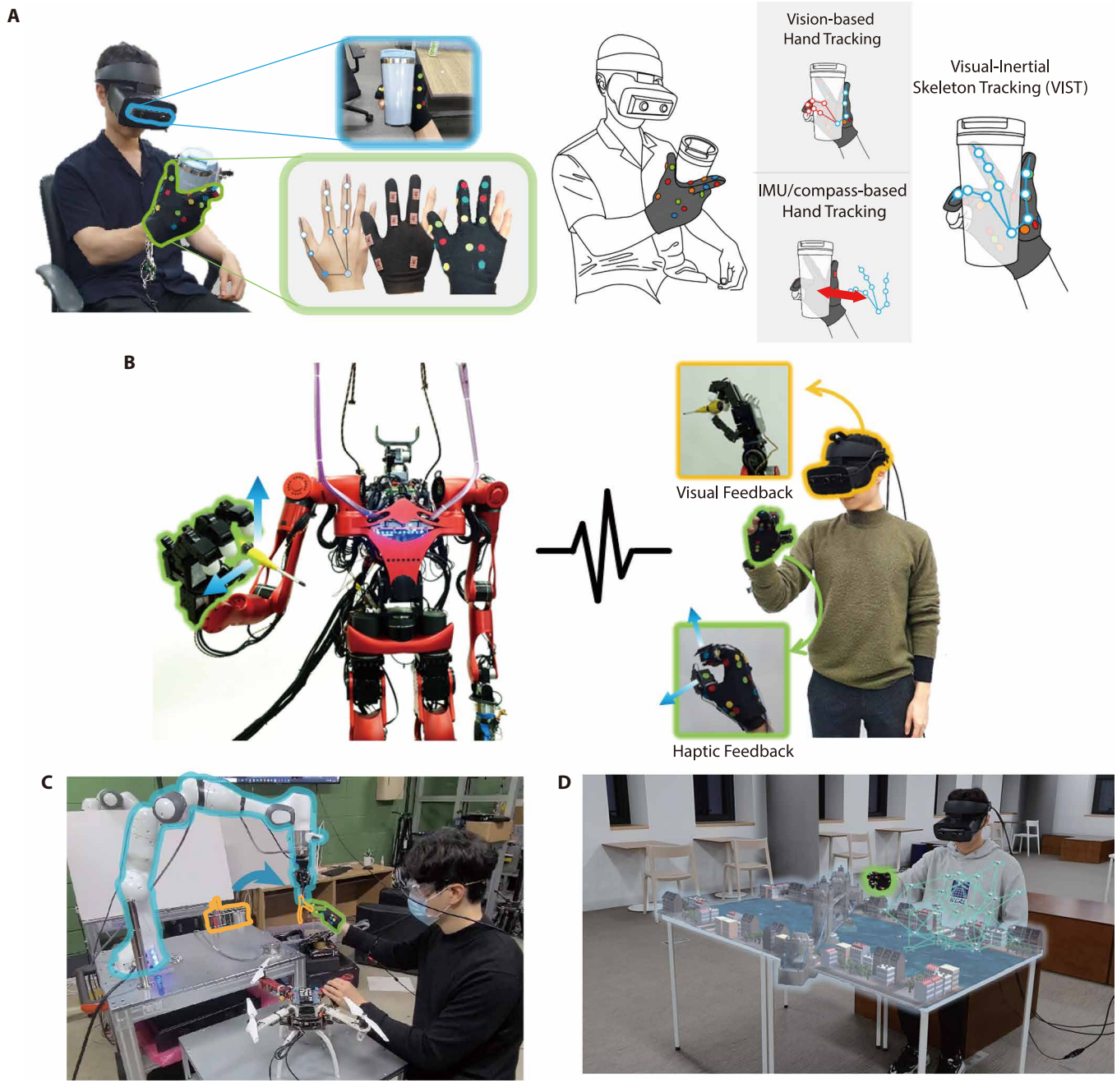
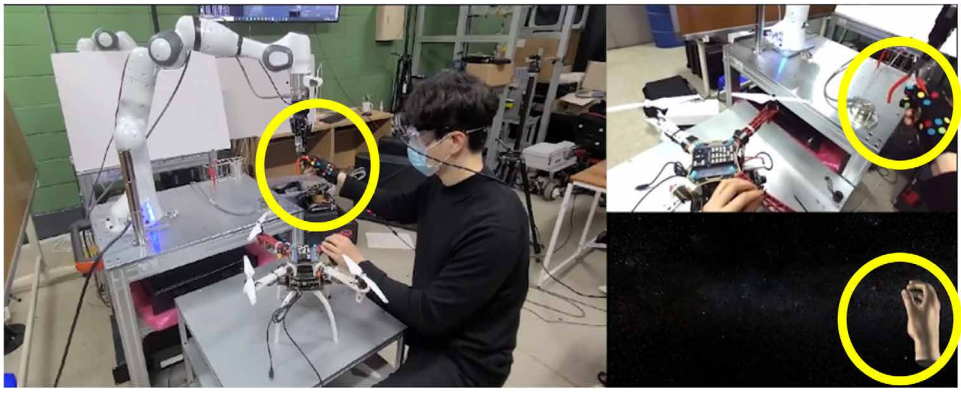



**Fig. 1. System configuration and possible applications of VIST.** (**A**) Hardware (i.e., sensor glove with IMUs/markers and stereo camera) and working principle of VIST. (**B**) Robot hand teleoperation (left image courtesy of DYROS/Seoul National University). (**C**) Collaborative robot interaction (Movie 1). (**D**) 3D drone swarm control (Movie 2).

while overcoming the fundamental limitations of the other methods as stated above (Fig. 1A and Movie 3). Specifically, we constructed a sensor glove integrated with seven IMUs and 37 visual markers (of four different colors). The glove is accompanied by a head-mounted stereo camera. We also developed a filtering-based visual-inertial hand tracking algorithm, with hand anatomical constraints taken into account along with the autocalibration of hand/sensor-related parameters (see Fig. 2). The stereo camera was chosen because of its availability and compatibility with VR/AR headsets; other vision sensors are equally applicable to our proposed VIST framework as explained in Materials and Methods.

One of the key innovations of our VIST framework is that we fuse the visual and inertial sensors in a tightly coupled (TC) manner. That is, we couple not only from visual to inertial [e.g., IMU drift correction (*40*, *41*)] but also from inertial to visual (e.g., IMU-aided correspondence search in the "Algorithms" section). TC fusion is key to coping with the peculiarity of hand tracking, where a number of skeletons (i.e., fingers) are moving fast and occlusions



**Movie 1. Collaborative robot interaction.** User can quickly and intuitively provide rich commands and cues to the robot using their hand and fingers, thereby making the interaction safer and fluidic as compared to the case of conventional pendant-based collaborative robot programming.

**Movie 2. 3D drone swarm control.** User can efficiently control the complex 3D swarm behavior by simply nudging their formation or quickly defining 3D virtual walls to avoid dangerous regions, all difficult when relying on conventional 2D tablet interface.

**Movie 3. Overview of VIST.** Summary of motivation, hardware construction, algorithm architecture, quantitative and qualitative experimental validation results, and possible applications of the proposed VIST framework.

occur in a small-size space (i.e., on the palm). The peculiarity requires us to use passive and anonymous markers (up to different colors), because the space is too small to accommodate tagged visual markers [e.g., AR markers (*42*) and VIVE trackers (*43*)], with their number also being as many as possible for robustness against the occlusions. With these anonymous visual markers, their correspondence search problem (an aspect that is central to the accurate vision processing) becomes very challenging. If it were not for this TC fusion, our VIST algorithm would fail the correspondence search, leading to unstable tracking (movie S1).

Several results have been proposed for the VIST (*40*–*43*), yet to our knowledge, very few adopt the TC approach and the majority rather rely only on the less accurate/robust loosely coupled (LC) approach [e.g., only IMU drift correction or just concatenate the two separate information (*30*)]. The LC fusion suffices for limb/torso skeleton tracking (*40*–*43*), because they all use only two or three visual markers (*41*, *40*) or tagged visual markers [e.g., AR markers (*42*) and VIVE trackers (*43*)]; thus, the correspondence search can be easily done even with the LC fusion. It is, however, not the case for the hand tracking as explained above. In contrast, our VIST framework brings in the TC fusion into the hand tracking, thereby achieving tracking accuracy and robustness at the same time.

Some of the important advantages of our VIST framework can be summarized as follows:

1) Superior tracking accuracy due to the TC visual-inertial fusion and the autocalibration as compared to other state-of-the-art approaches (see the "Quantitative evaluation" section in Results).

2) Robustness against occlusions, visually complex/changing environments, and ambient lighting (movies S2 to S6).

3) Robustness against electromagnetic interference and ambiguous mechanical contacts, thus enabling object manipulation and the device to be worn (movies S4, S7, and S8).

4) Convenience of real-time calibration/autocalibration of anatomical/glove kinematic parameters integrated into the VIST algorithm.

5) Ruggedness (e.g., washable; movie S9), affordability [e.g., total glove material cost ≈$100 (fig. S1) with the cameras and computing of the head-mounted display (HMD) possible to use], and wearability [e.g., light weight (52 to 55 g)].

Our VIST framework may also be used to collect human data that could be used to train reinforcement learning algorithms to learn robotic object manipulation (*44*) or as a tracking module for the feedback control of soft robotic hand prostheses (*45*, *46*). Our VIST framework can further be used for robotic systems with limbs, for which typical proprioceptive sensors are difficult to deploy (e.g., very thin tendon-driven robots or soft multilegged robots with frequent whole-body contacts).

## RESULTS

### Human hand and sensor glove models

The human hand is modeled as a segment joint skeleton model (Fig. 2A) (*47*), where the types of joints are determined according to

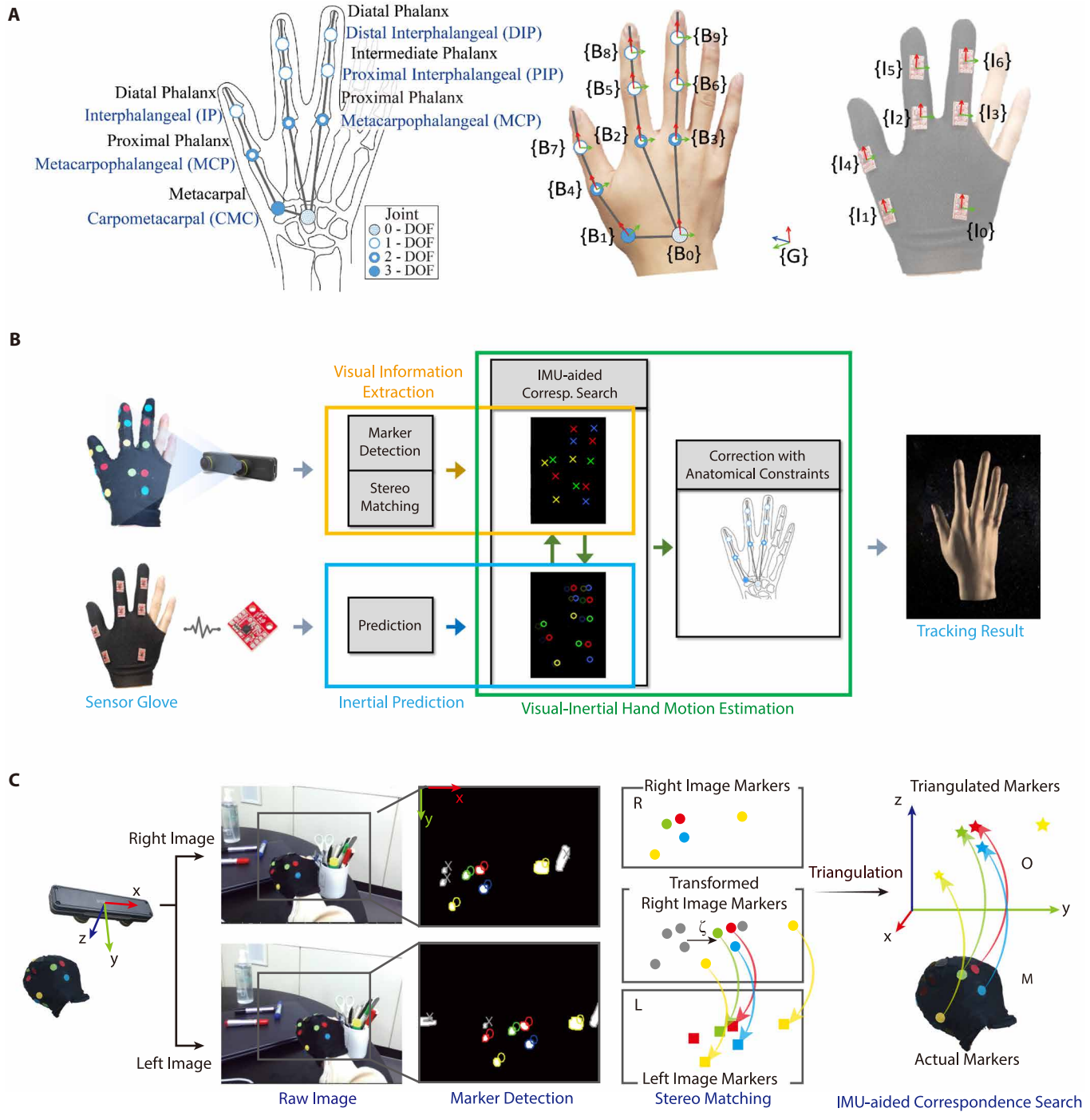



**Fig. 2. Modeling and processes of VIST.** (**A**) Hand and glove models with coordinate frames. (**B**) Visual-inertial fusion pipeline. (**C**) Visual information extraction process with marker detection, stereo matching, and IMU-aided correspondence search.

their anatomical structures. This segment joint model is adopted in many model-based hand tracking systems with vision, IMU/compass, or soft sensors (*22*, *23*, *48*, *49*). Here, we choose the target tracking segments to be the dorsum of the hand and the three (thumb, index, and middle) fingers; these segments play a key role in our daily activities and influence the motions of the ring/little fingers (*50*). We also assume musculoskeletal dependencies [i.e., synergy (*22*, *23*, *51*)] to estimate the angle of the interphalangeal (IP) joint from the metacarpophalangeal (MCP) joint for the thumb, and that of the distal IP (DIP) joints from the proximal IP (PIP) joints for the index and middle fingers. Because our VIST algorithm is applicable to any skeletal tracking with segments and joints, its extension to the ring/little fingers or to the case of no synergy is straightforward.

To obtain the visual and inertial information of the target tracking segments, we design a sensor glove comprising two layers: an inner glove layer with seven IMUs (on the dorsum of the hand,

metacarpal/proximal phalanges of the thumb, and proximal/intermediate phalanges of the index/middle fingers) and an outer glove layer with 37 visual markers [fabric blobs with four different colors (red, blue, green, and yellow)] (see fig. S1 and Materials and Methods). We can then define two types of coordinate frames: the coordinate frame of the $i$th hand segment $\{B_i\}$ ($i$ = 0,1,2, …,9) and the coordinate frame of the $j$th IMU $\{I_j\}$ ($j$ = 0,1,2, …,6), where the IMU frame index $j$ is the same as that of its corresponding hand segment index $i$ whenever relevant (Fig. 2A).

The origin of $\{B_j\}$ is attached to its parental joint, each axis of which is along the axes of flexion/extension ($y$ axis), abduction/adduction ($z$ axis), and twisting ($x$ axis), whereas each $\{I_j\}$ is attached to its IMUs, whose pose is not necessarily matched with corresponding $\{B_j\}$. Thus, many IMU-based tracking systems attempt to align $\{I_j\}$ with $\{B_j\}$ when attaching IMUs or require calibration before the operation through a sequence of indicated postures (*52*, *53*, *54*). However, misalignment error is inevitable when attaching IMUs. Moreover, calibration is often not precise because there is some human error in the process of capturing the indicated postures. Our VIST algorithm, in contrast, can deal with such errors in real time through autocalibration (see the next section), thereby substantially improving tracking performance.

## Algorithms

Using the models in Fig. 2 as explained above, we design our VIST algorithm comprising the following two parts (Fig. 2B): visual information extraction, which robustly obtains the 3D positional observations of the many/anonymous visual markers via the stereo camera and TC fusion with IMU information, and visual-inertial hand motion estimation, which estimates the hand motion by fusing the IMU information with the extracted visual information and the hand anatomical constraints.

### *Visual information extraction*

The visual information extraction process comprises three subprocesses (Fig. 2C): marker detection in raw images, left-right stereo matching, and IMU-aided correspondence search.

*Marker detection in raw images.* To detect the visual markers (i.e., color blobs) in the raw stereo images, we use the following two requirements standard in the field of computer vision: (i) Hue-saturation values (HSVs) requirement, that is, we extract only the visual blobs having HSV within predefined intervals of the blob colors, and (ii) shape requirements (i.e., size, convexity, and circularity), that is, we extract only the visual blobs with reasonable size and shape based on their real size and the distance from the camera. The centroids of the blobs satisfying both the HSV and the shape requirements are then determined as the 2D pixel observation sets of the markers, i.e., $\mathcal{R} = \{r_1, r_2, \ldots r_{N_R}\} \in \mathfrak{R}^{2N_R}$ for the right image and $\mathcal{L} = \{l_1, l_2, \ldots l_{N_L}\} \in \mathfrak{R}^{2N_L}$ for the left image, respectively.

*Left-right stereo matching.* We obtain the 3D positional observations of the visual markers by matching/triangulating each pair of points in the left and right observation sets, $\mathcal{L}$ and $\mathcal{R}$. For this, we use a coherent point drift (CPD) algorithm (*55*), a classical point set registration method (see note S1). Specifically, to match the two point sets, the CPD algorithm represents one point set as a Gaussian mixture model (GMM) and determines the solution (i.e., transformation and correspondences between the two sets) using an expectation-maximization (EM) algorithm to maximize the product of the correspondence probability and the transformed GMM probability of the other set. In stereo matching, without loss of generality, we represent $\mathcal{R}$ as a GMM and define a transformation parameter $\zeta \in \mathfrak{R}$ to represent the 1-DOF horizontal parallax between the two sets, neglecting vertical transformation with the assumption that the raw stereo images have been rectified. The transformation parameter $\zeta$ is then obtained by using the EM algorithm as in (*55*). We choose the CPD algorithm here due to its simplicity in the rigid point set registration. Other methods [e.g., (*56*, *57*)] are available depending on the matching complexity.

Once $\zeta$ is determined, each left point $l_i \in \mathcal{L}$ has a matched candidate $r_{j,\min} \in \mathcal{R}$, which is the closest right point when transformed by $\zeta$. Because it is possible that blobs are observed only from one (left or right) camera, we define two additional conditions based on Euclidean distance to identify such outliers (note S2). When $l_i$ and $r_{j,\min}$ satisfy these two conditions simultaneously, the points are matched; otherwise, $l_i$ is identified as an outlier. Then, the matched points are triangulated using the mechanical specifications (i.e., focal length and baseline) of the adopted camera, and the 3D positional observations of the markers, $\mathcal{O} = \{o_1, o_2, \ldots, o_{N_O}\} \in \mathfrak{R}^3$, are constructed, where $N_O$ is the number of the matched points from the stereo images.

*IMU-aided correspondence search.* This process aims to find the correspondence of the set of the stereo-matched markers $\mathcal{O}$ to the set of the IMU-predicted positions of the visual markers (via (9)), $\mathcal{M} = \{m_1, m_2, \ldots m_{N_M}\} \in \mathfrak{R}^3$, where $N_M$ = 37 (i.e., number of all the visual markers attached to the glove). We again apply the CPD algorithm to match $\mathcal{O}$ and $\mathcal{M}$, by defining $\mathcal{M}$ as a GMM and finding the transformation parameter $\eta \in \mathfrak{R}^3$, which only represents the 3-DOF translation between the two sets, because the rotation of $\mathcal{M}$ can be updated fairly precisely with the gyroscope over a short period of time (*22*). The GMM likelihood function of the set $\mathcal{O}$ is then defined as

$$\mathrm{p}\left(\mathcal{O} \mid \mathcal{M}, \eta\right) = \prod_{h=1}^{4} \prod_{i=1}^{N_{O,h}} \mathrm{p}\left(o_{i,h} \mid \mathcal{M}_h, \eta\right) \quad (1)$$

where $N_{O,h}$ is the number of marker observations for each hue $h \in \{1,2,3,4\}$ (i.e., red, blue, green, and yellow).

Because $\mathcal{M}$ is represented as the GMM centroids, the probability of each marker observation $o_{i,h} \in \mathcal{O}$ is given by

$$\mathrm{p}(o_{i,h} \mid \mathcal{M}_h, \eta) = (1 - w_c) \sum_{j \in \mathcal{M}_h} \mathrm{p}(m_j)\, \mathrm{p}(o_{i,h} \mid m_j, \eta) + w_c / N_{O,h} \quad (2)$$

$$\mathrm{p}\left(o_{i,h} \mid m_j, \eta\right) \sim \mathcal{N}\left(\mathrm{T}(m_j, \eta), \Sigma_c\right) \quad (3)$$

where $\mathrm{T}(m_j, \eta) \in \mathfrak{R}^3$ is the transformation of the point $m_j \in \mathcal{M}$ using the parameter $\eta \in \mathfrak{R}^3$, $w_c \in [0,1]$ is the parameter determining the outlier ratio of the correspondence search, and $\Sigma_c \in \mathfrak{R}^{3\times3}$ is the covariance matrix of observation noises of the marker triangulation (fig. S2), which is obtained by pilot tests.

It is typical that only less than one-third of all the (37) makers survive the stereo matching [i.e., average($N_O$) = 10.93; see Fig. 3E]. Furthermore, because those markers in $\mathcal{O}$ are anonymous (up to different colors) and the 3D motion of each finger is fast, if we attempt the correspondence search only with the vision information via CPD (i.e., $\widehat{\mathcal{O}}_{t-1} \rightarrow \mathcal{O}_t$, where $\widehat{\mathcal{O}}_{t-1} \in \mathfrak{R}^{3\times37}$ is the computed positions of all the markers based on the hand motion estimated at the previous time $t-1$, whereas $\mathcal{O}_t \in \mathfrak{R}^{3N_O}$ is the stereo-matched markers at the current time $t$), the search very often ends up being an outlier and the hand tracking becomes unstable (movie S1). To circumvent this, we compute the prior $\mathrm{p}(m_j)$ of each marker from the latest IMU-predicted hand pose ($m_j \in \mathcal{M}$) and perform the

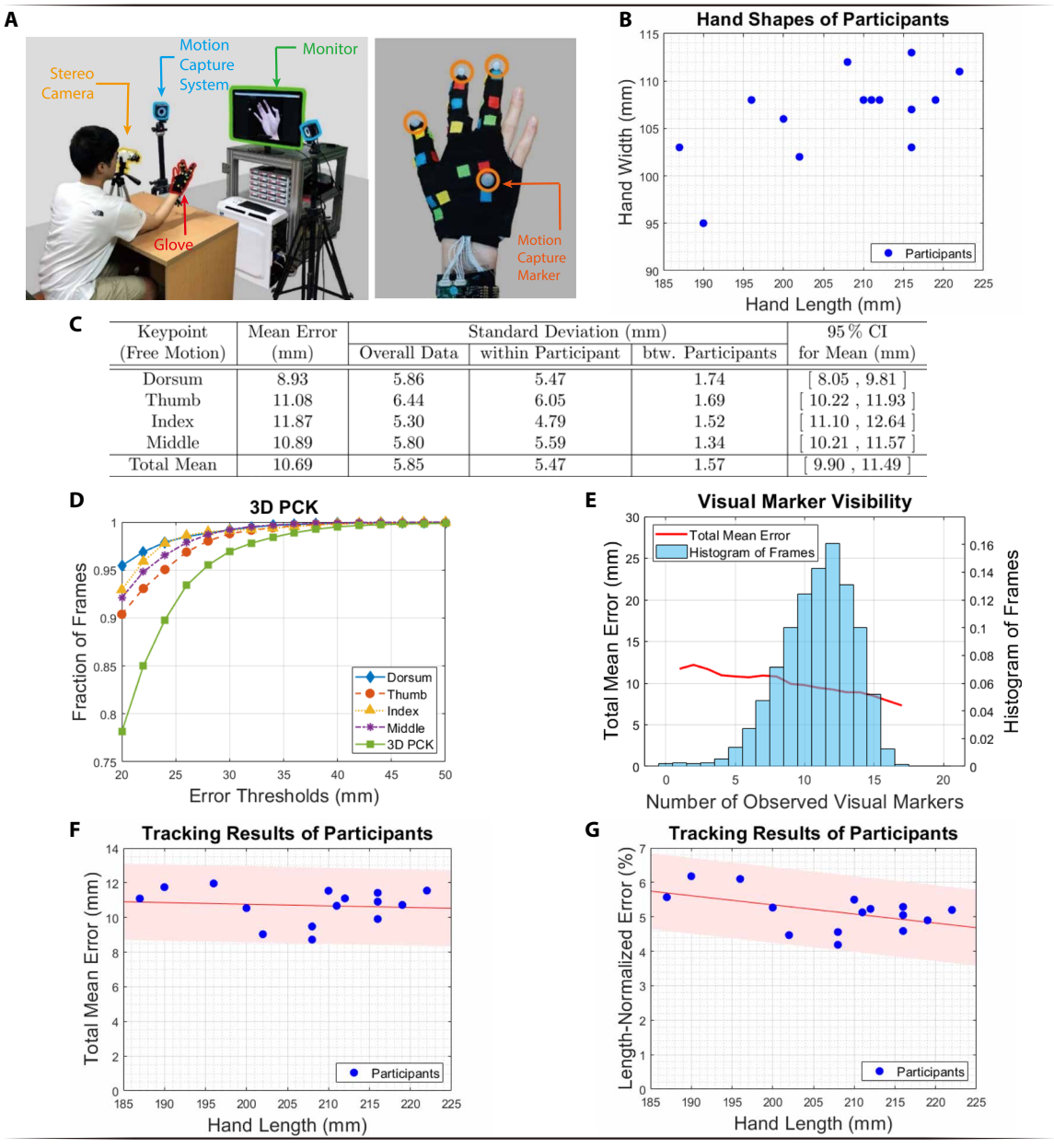
correspondence search $\mathcal{M}_t \rightarrow \mathcal{O}_t$. This renders our VIST algorithm to be TC fusion, with its accuracy and robustness markedly improved. More precisely, we consider the following factors: camera-facing factor and FOV factor (fig. S3).

The camera-facing factor considers a marker as difficult to observe when its normal vector is in the direction opposite to the camera center. The observation probability $p_n(m_j)$ for the camera-facing factor is defined as

$$p_n(m_j) = \begin{cases} 1, & \text{for } \alpha_{m_j} \leq \alpha_{\min} \\ (\alpha_{\max} - \alpha_{m_j})/(\alpha_{\max} - \alpha_{\min}), & \text{for } \alpha_{\min} < \alpha_{m_j} \leq \alpha_{\max} \\ 0, & \text{for } \alpha_{\max} < \alpha_{m_j} \end{cases} \quad (4)$$

where $(\alpha_{\min}, \alpha_{\max})$ is the visible angular range of the adopted marker type and $\alpha_{m_j}$ is the angle between the normal vector and the camera ray (i.e., line from camera center to marker) of the marker $m_j$.

The FOV factor excludes any visual markers outside the FOV from the correspondence search, thus enhancing the tracking robustness when the hand is partially observed around the edge of the FOV. Given the IMU-predicted marker position $m_j \in \Re^3$ and its covariance matrix $\Sigma_j \in \Re^{3\times3}$ from the estimator, we compute the image plane–projected marker position $m_j^* \in \Re^2$ and its covariance matrix $\Sigma_j^* \in \Re^{2\times2}$. Near the edge of the FOV, the observation probability $p_f(m_j)$ for the FOV factor is given by integrating the area inside the FOV of the Gaussian distribution of $m_j^*$ such that




| Keypoint (Free Motion) | Mean Error (mm) | Standard Deviation (mm) | | | 95% CI for Mean (mm) |
|---|---|---|---|---|---|
| | | Overall Data | within Participant | btw. Participants | |
| Dorsum | 8.93 | 5.86 | 5.47 | 1.74 | [ 8.05 , 9.81 ] |
| Thumb | 11.08 | 6.44 | 6.05 | 1.69 | [ 10.22 , 11.93 ] |
| Index | 11.87 | 5.30 | 4.79 | 1.52 | [ 11.10 , 12.64 ] |
| Middle | 10.89 | 5.80 | 5.59 | 1.34 | [ 10.21 , 11.57 ] |
| Total Mean | 10.69 | 5.85 | 5.47 | 1.57 | [ 9.90 , 11.49 ] |

**Fig. 3. Test setup and results of quantitative evaluation of tracking in free motion.** (**A**) Quantitative evaluation test setup. (**B**) Participant hand size distribution (see also fig. S6). (**C**) Statistical analysis results of the four keypoints (see also fig. S7). (**D**) 3D-PCK analysis. (**E**) Histogram of frames and mean errors versus the number of observed markers. (**F**) Absolute mean error. (**G**) Length-normalized error distributions versus the hand length.

$$p_f(m_j) = \int_{-\infty}^{d_j} \left( e^{-(x^2)/(2\sigma_j^{*2})} \right) / \left( \sqrt{2\pi\hat{\sigma}_j^{*2}} \right) dx \quad (5)$$

where $d_j \in \Re$ is the distance between $m_j^*$ and the nearest edge of the FOV and $\sigma_j^*$ is an element of $\Sigma_j^*$ corresponding to the direction of $d_j$.

Using the camera-facing factor (Eq. 4) and the FOV factor (Eq. 5), the final observation probability of the marker $m_j$ is the product of these two probabilities for both cameras, that is,

$$p(m_j) = \prod_{c=R,L} p_{n,c}(m_j)\, p_{f,c}(m_j) \quad (6)$$

where $c$ is the index of the right or left camera. We also eliminate a marker $m_j$ occluded by other segments in the point matching [i.e., $p(m_j)$ becomes 0], if the image plane–projected marker position $m_j^*$ is inside the projected shape of another segment and the marker is farther away than the segment. This prior probability of each marker $p(m_j)$ is then used to compute the posterior probability (Eq. 2) for the correspondence search.

Once the transformation parameter $\eta \in \Re^3$ for Eq. 1 is computed using the EM algorithm, an observation $o_i \in \mathcal{O}$ is assigned to the IMU-predicted marker $m_j \in \mathcal{M}$ with the maximum matching probability $p(m_j \mid o_i)$ from all the markers $\mathcal{M} = \{m_1, m_2, \ldots m_{N_M}\}$ as described in note S3. However, this may still match a single $m_j$ to multiple observations $o_i$. To reject this duplicated match, as in the stereo matching, similar to that in (*58*), we introduce a threshold condition based on Euclidean distance to identify outliers. Then, finally, we attain the IMU-aided correspondence search: $\mathcal{M} \supset \hat{\mathcal{Z}} \rightarrow \mathcal{Z} \subset \mathcal{O}$, where $\hat{\mathcal{Z}}$ and $\mathcal{Z}$ are the sets of the corresponding markers in $\mathcal{M}$ and $\mathcal{O}$ with the same dimension and will be used for the extended Kalman filter (EKF) update (Eq. 10).

***Visual-inertial hand motion estimation***

For the visual-inertial sensor fusion with a large number of states at a rate faster than the sensor sampling rate, we deploy the EKF, which is the most common estimator for nonlinear systems and exhibits reasonable performance with limited computation load (fig. S4). This EKF consists of three subprocesses: prediction with IMU information, correction with visual information, and correction with anatomical constraints.

***EKF states for hand tracking and autocalibration***

We define the EKF states for each segment of the hand (total seven segments; see Fig. 2A) as

$$x := [x_s; x_p] \in \Re^{23} \quad (7)$$

where $x_s := \left[p_{G,i}^G; v_{G,i}^G; q_{G,i}^G; b_g; b_a\right] \in \Re^{16}$ is the motion-related state of the segment, which includes the position, velocity, unit quaternion of the the IMU coordinate frame $\{I\}$ in the global coordinate frame $\{G\}$, and the IMU biases adopting the model of (*59*). We also define the state for the autocalibration of hand/sensor-related kinematic parameters such that

$$x_p := \left[\lambda_B; q_{I,B}^I\right] \in \Re^7 \quad (8)$$

where $\lambda_B \in \Re^3$ is the scale factor of the attached segment $\{B\}$, which is dependent on the user hand size, and $q_{I,B}^I \in \Re^4$ is the quaternion for misalignment between $\{I\}$ and $\{B\}$, which is dependent on the user hand shape and may also change for each fitting (fig. S5).

Inclusion of this autocalibration is one of the key strengths of our VIST framework. Vision-based tracking systems with machine-learning techniques are well known for their fragility and loss of performance for hand shapes/configurations outside their training sets, whereas those based on IMUs/compass or soft sensors rely on the assumption that users can/will precisely reproduce all the indicated poses, which is not true in practice and results in typically less accurate calibration and fingertip position tracking. In contrast, due to the real-time calibrations/autocalibrations of $x_p$ using the visual-inertial fusion, our proposed VIST framework can substantially improve tracking accuracy and user convenience as compared to other systems.

***Prediction with IMU information***

In the prediction step, the nominal state and its covariance matrix are predicted with the IMU information using the following kinematic model for each hand segment

$$\hat{\dot{x}} = f(\hat{x}, a_m, w_m) \quad (9)$$

where $a_m \in \Re^3$ and $w_m \in \Re^3$ are respectively the accelerometer and gyroscope data of the IMU, and $\hat{x} \in \Re^{23}$ is the predicted state with the IMU information. This kinematic model (Eq. 9) is derived from note S4. The linearized error state propagation model in Eq. 9 can then be obtained as $\tilde{\dot{x}} = F\tilde{x} + Gn$, where $\tilde{x} \in \Re^{21}$ is the error state, $F \in \Re^{21\times21}$ is the error state transition matrix containing $(a_m, w_m)$ from the IMU, $G \in \Re^{21\times18}$ is the input noise matrix, and $n \in \Re^{18}$ is the concatenated noise vectors. The dimension difference between $\hat{x}$ and $\tilde{x}$ is due to the error state computation of the quaternion quantities. Expressions of $F$, $G$, and $n$ are provided in note S5. Because the IMU measurements are sampled at the IMU sampling rate, we discretize the continuous time prediction model for the VIST implementation.

***Correction with visual information***

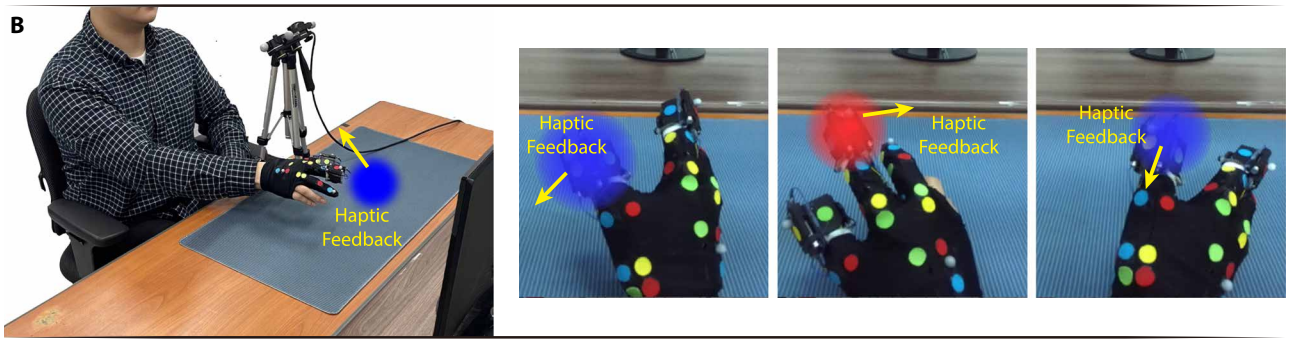
The IMU-predicted motion of each segment is generally inaccurate because of the lack of compass information, sensor noise, and uncalibrated parameters (i.e., $b_g, b_a, \lambda_B, q_{I,B}^I$). Thus, we correct this IMU-predicted hand segment motion and also the uncalibrated parameters using the correspondence-matched marker measurements $\mathcal{Z} \subset \mathcal{O}$. Specifically, we use the linearized error model of the measurement equation such that

$$\tilde{z}_{m_j} = z_{m_j} - \hat{z}_{m_j} \simeq H_{m_j}\tilde{x} + n_z \quad (10)$$

where $z_{m_j} \in \Re^3$ is the measurement of $o_j \in \mathcal{Z} \subset \mathcal{O}$ with respect to the global coordinate frame, $\hat{z}_{m_j} = h(\hat{x}) \in \Re^3$ is that of the IMU-predicted marker $m_j \in \hat{\mathcal{Z}} \subset \mathcal{M}$, which corresponds to $o_j$ at the current time, and $n_z \in \Re^3$ is the noise from triangulated marker measurements, which is modeled as a zero mean white Gaussian and follows the covariance matrix $\Sigma_c \in \Re^{3\times3}$ in Eq. 3. The observation matrix $H_{m_j} \in \Re^{3\times21}$ is the Jacobian of the measurement equation $h(\tilde{x})$ with respect to $\tilde{x}$. Detailed derivations of these measurement equations are explained in note S6. For delay-free estimations even in the case of fast hand motions, similar to (*60*), we use a ring buffer to synchronize the current IMU data with the delayed visual data (about tens of milliseconds delay).

***Correction with anatomical constraints***

Although we estimate the motion of each segment independently as a free rigid body as stated above, their motions are not independent but rather anatomically correlated. We thus formulate some anatomical constraints of the human hand as the measurement equations for the


C

| Keypoint (Object Interaction) | Mean Error (mm) | Standard Deviation (mm) | | | 95% CI for Mean (mm) |
|---|---|---|---|---|---|
| | | Overall Data | within Participant | btw. Participants | |
| Dorsum | 9.85 | 7.06 | 6.69 | 1.51 | [ 9.08 , 10.61 ] |
| Thumb | 12.29 | 7.37 | 7.02 | 1.88 | [ 11.34 , 13.24 ] |
| Index | 14.04 | 6.30 | 6.00 | 1.54 | [ 13.26 , 14.82 ] |
| Middle | 14.52 | 5.78 | 5.47 | 1.71 | [ 13.66 , 15.39 ] |
| Total Mean | 12.68 | 6.63 | 6.30 | 1.66 | [ 11.84 , 13.52 ] |

D

| Keypoint (Wearing CHD) | Mean Error (mm) | Standard Deviation (mm) | | | 95% CI for Mean (mm) |
|---|---|---|---|---|---|
| | | Overall Data | within Participant | btw. Participants | |
| Dorsum | 9.14 | 4.15 | 3.76 | 1.64 | [ 8.31 , 9.97 ] |
| Thumb | 10.82 | 5.83 | 5.61 | 1.52 | [ 10.05 , 11.59 ] |
| Index | 12.22 | 6.04 | 5.73 | 1.41 | [ 11.50 , 12.93 ] |
| Middle | 11.38 | 6.49 | 6.17 | 1.56 | [ 10.59 , 12.17 ] |
| Total Mean | 10.89 | 5.63 | 5.32 | 1.53 | [ 10.11 , 11.66 ] |

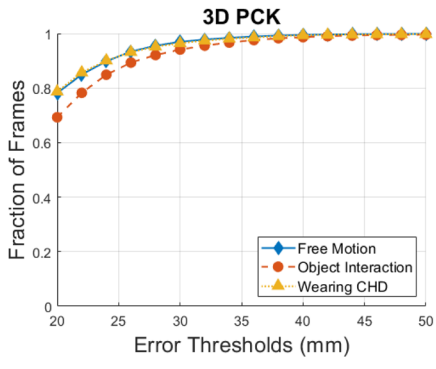



**Fig. 4. Test setup and results of quantitative evaluation of tracking with object interaction and wearable device.** Test setups for object interaction (**A**) and wearing CHD (**B**) with their statistical analysis (**C** and **D**) and 3D-PCK analysis (**E**).

EKF correction. We first define the positional constraint to force anatomically adjacent segments to be connected at their pivot joint (e.g., intermediate and proximal phalanges connected at the PIP joint) by enforcing their global positions to be the same. A total of six such positional constraints are applied [carpometacarpal (CMC)/MCP joints for the thumb and MCP/PIP joints for the index and middle fingers]. We also define the rotational constraint (e.g., PIP joint cannot twist about the $x$ axis). A total of seven rotational constraints are applied (no $x$-axis rotation of MCP joints for the three fingers and no $x$/$z$-axis rotations of the PIP joints for the index and middle fingers) following the adopted anatomical model (Fig. 2A). The observation matrix for both the positional and rotational constraints is derived from note S7.

## Quantitative evaluation

Here, we quantitatively evaluate the performance and robustness of our proposed VIST framework in the cases of free motion, object interactions, and wearing fingertip cutaneous haptic devices (CHDs). For this, we attach motion capture (MOCAP) markers to four keypoints of the glove (three at the fingertips and one on the hand dorsum; see Fig. 3A), because (i) the MOCAP system cannot track robustly more than four markers, since they are anonymous and moving within a small size space, which is the very motivation of our VIST framework, and (ii) the fingertips typically exhibit larger tracking error than other parts of the hands [e.g., MCP, DIP, or PIP (*10*, *12*)] with their accurate tracking crucial for some important applications (e.g., telepicking via pinching and fingertip touch in VR). As shown in Fig. 3A, each participant was instructed to sit in front of a table surrounded by the MOCAP cameras and to duplicate a hand configuration randomly displayed in the monitor. Fifteen participants were recruited for each experiment. Details on the experimental setup and procedure are explained in Materials and Methods and movie S10.

### *Hand tracking in free motion*

We first evaluate our proposed VIST framework for the case of free hand motion as frequently adopted to quantitatively evaluate

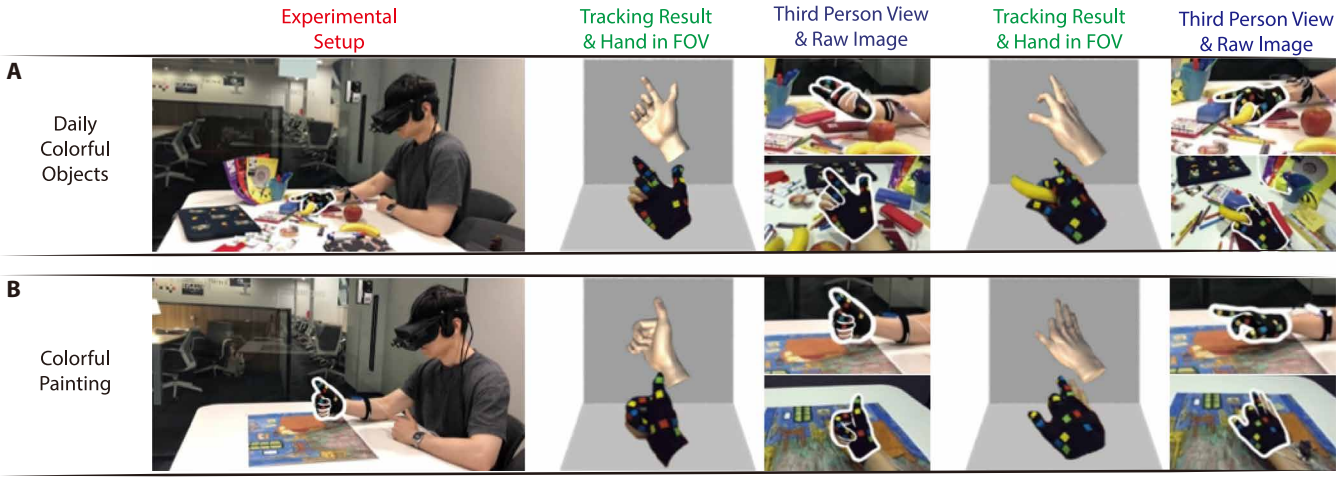



**Fig. 5. Qualitative evaluations for visually complex backgrounds.** (**A**) Colorful objects in the background or manipulated by hands. (**B**) Colorful painting background.

baseline tracking performance [e.g., (*10*, *12*, *13*)]. Seated in the setup of Fig. 3A, each participant was instructed to copy the hand posture displayed on the monitor, which was randomly chosen from the large (tens of thousands) compounded hand image datasets of (*10*, *61*). The hand image was switched to the next one every 3 s, and each trial lasted 3 min (i.e., a total of 60 hand postures) for each participant.

Statistical measures for each of the four keypoint are provided in Fig. 3C, with their total means across the four keypoints listed in the last row. There, we can see (i) good tracking with the total mean error of 10.69 mm and SD of 5.85 mm; (ii) robustness against participant variability (see Fig. 3B), with the within-participant SD (i.e., mean of SD of each participant) similar to the total SD and the between-participant SD (i.e., SD of mean error of each participant) fairly small of 1.57 mm (see also fig. S7); and (iii) good statistical confidence with a narrow mean error 95% CI (confidence interval) of [9.90 mm, 11.49 mm]. We also compute the 3D percentage of correct keypoints (3D-PCK), i.e., the fraction of the frames, for which the worst tracking error of all the four keypoints is within a given error bound (*62*). This PCK is a popular evaluation measure (*10*, *12*, *13*, *19*). From Fig. 3D, we can see good tracking performance with 3D-PCK@20mm ≈ 78% and 3D-PCK@35mm≈ 99%.

The mean errors and the histogram of frames with respect to the number of observed markers are shown in Fig. 3E, where we can see the good robustness of our VIST against occlusions with its mean error only mildly increasing even with fairly few visible markers. This is due to VIST's ability to exploit the IMU information and contrasts with typical pure vision–based approaches (*10*, *12*, *19*), which typically suffer from sharp tracking loss with occlusions. Note also that the histogram shape resembles a normal distribution; that is, the displayed hand images are not biased and the evaluation here is applicable to general/diverse hand postures. We also present plots of the mean error and the hand length–normalized error with respect to the participant hand lengths in Fig. 3 (F and G). There, we can see that our VIST tracking, because of its utilization of vision information and autocalibration of hand parameters, is insensitive to hand size variability and actually enhances the length-normalized error. This is in contrast to the IMU/compass and soft wearable tracking approaches (well known for the error amplification with respect to their size) (*21*, *26*) as well as to the pure vision–based approaches (well known for the issue of generalization outside of their training sets) (*19*).

***Hand tracking with object interaction and wearable device***

We quantitatively evaluate our VIST framework for the two scenarios, object interactions and wearing fingertip CHDs (*22*) (see also Fig. 8), to show its robustness against severe occlusions, magnetic interference, and mechanical contacts. For the object interaction, we construct our own image datasets of eight daily objects (fig. S8), with 100 images of different hand postures for each object (i.e., a total of 800 images). For each participant, one of the eight objects was randomly selected, and a 10-s period was given for them to hold the chosen object. Each participant was then instructed to copy the object-hand posture displayed on the monitor every 3 s. This procedure was repeated for another two objects, making up a total 2-min trial for each participant (Fig. 4A). For the evaluation with CHD, a translucent virtual sphere was rendered in blue or red color and with its location/size randomly varying. This sphere was then projected into the left image of the (standing) stereo camera (Fig. 4B), and this left image was displayed on the monitor. Each participant was then asked to move their index (or thumb, respectively) fingertip to the center of the blue (or red, respectively) sphere with penetration proportional normal force feedback from CHD. A total of 40 virtual spheres were rendered, each lasting 3 s, constituting a 2-min trial for each participant.

Statistical measures of the four keypoints with the object interaction and the wearing CHD are summarized in Fig. 4 (C and D) (see also fig. S7) along with their 3D-PCK in Fig. 4E. There, similar for Fig. 3, we can see that our VIST framework still exhibits good tracking (e.g., mean errors of 12.68/10.89 mm with 3D-PCK@20mm ≈ 69/79% and 3D-PCK@35mm ≈ 96/98%), robustness against participant variability (e.g., within-participant SDs ≈ total SDs, between-participant SDs of 1.66/1.53 mm), and good statistical confidence (with 95% CIs of [11.84 mm, 13.52 mm]/[10.11 mm, 11.66 mm]). Although these measures are a bit deteriorated from that of the free motion in Fig. 3, the efficacy of our VIST framework could still be asserted in terms of human perception. More precisely, the work of (*63*) shows that humans cannot detect index fingertip tracking errors in VR under 50 mm, whereas the work of (*64*) shows that humans cannot discriminate index finger joint angle error under 1.7° based on proprioception. However, for our VIST framework, 3D-PCK@50mm

is greater than or equal to 99% for both Figs. 3 and 4, whereas its angle tracking error would be less than 1.57° of (*22*), given that the fingertip tracking error of (*22*) is larger (i.e., mean error about 29 mm) than our VIST framework. This suggests that our VIST framework would likely allow users to perceive their rendered hands accurately following their real hands.

We also note that our VIST framework outperforms some state-of-the-art vision-based algorithms: (i) for the free motion tracking, our VIST algorithm exhibits a mean error of 10.69 mm and 3D-PCK@20/35mm ≈ 78/99%, whereas the work of (*10*) had a mean error of about 50 and 3D-PCK@20/35mm ≈ 44/65% [even if evaluated against only a subset (*61*) of our datasets in (*10*)], and the winners of the latest hand tracking challenge (*19*) had a mean error of 13.66 mm (*65*) and 3D-PCK@20/35mm ≈ 24/67% (*66*) [even if the hand bounding box is given in (*65*, *66*)], and (ii) for the object interactions, our VIST algorithm provides a mean error of 12.68 mm and 3D-PCK@20/40mm ≈ 69/97%, whereas the winner (*11*) of the challenge (*19*) provided a mean error of 24.74 mm and 3D-PCK@20/40mm ≈ 0/27% [even if the wrist ground truth position is given in (*11*)].

Although these indirect comparisons with (*10*, *65*, *66*) are still indicative enough of enhanced performance of our VIST, we also implement the algorithm of (*9*) and directly apply it to the same object-hand image datasets and the same procedure with wearing the CHD. Here, we choose (*9*) as it is one of the most advanced vision-based algorithms so far, while others are not as generalizable [e.g., bone lengths required (*10*)] or not open to public [e.g., (*11*)]. We then found that this algorithm (*9*) cannot maintain tracking stability during the object interactions or wearing CHD (fig. S9), which is expected, because the issue of occlusions is a well-known problem for any pure vision–based algorithms. Hand tracking with objects/devices is, of course, difficult for other methodologies as well, and for this, we apply the IMU/compass-based tracking of (*22*) to the datasets/procedure of this section and found that, with hand drill, portable fan, earphone case, and CHD, all containing ferromagnetic materials or internal current, the tracking becomes unstable with some finger joints excessively twisted (fig. S10). In contrast, due to its opportunistically exploiting vision and IMU informations, our VIST can maintain not only tracking stability but also its accuracy as shown in Figs. 3 and 4.

## Qualitative evaluation in real-world scenarios

Here, we perform qualitative evaluation of our VIST framework for some challenging real-world scenarios that outperform existing hand tracking methodologies.

### *Visually challenging background*

It is challenging for vision-based systems to track the human hand on backgrounds with similar appearances/colors (*10*, *13*, *67*). To evaluate robustness against such situations, we designed qualitative experiments with colorful objects (magazines, fruits, and stationery) and a painting (*Bedroom in Arles*) in the background with visually similar colors/patterns to the glove markers. As shown in Fig. 5 and movie S5, despite the visually adversarial objects/backgrounds, our VIST can robustly track the hand motion even when it interacts with daily objects (bananas and scissors). This is because our VIST algorithm accurately detects visual markers from the background using the HSV and shape requirements simultaneously. Moreover, through the IMU-aided correspondence search, it can robustly match the marker observations with the true anonymous markers on the glove while effectively eliminating outliers, thereby exhibiting this stable tracking even with the visually complex objects/backgrounds.

Hand tracking outdoors is a difficult problem for existing tracking systems. RGB-D vision–based hand tracking is generally not suitable for outdoor use, because sunlight can interfere with structured infrared (IR) rays (*68*, *69*). RGB vision–based tracking does not work well in outdoor settings either, because their training sets are typically acquired indoors (*13*, *16*, *61*). Typical IMU/compass or soft sensor wearable hand tracking (*22*, *23*, *27*) also suffers from the same problem outdoors, because they also typically require IR-based



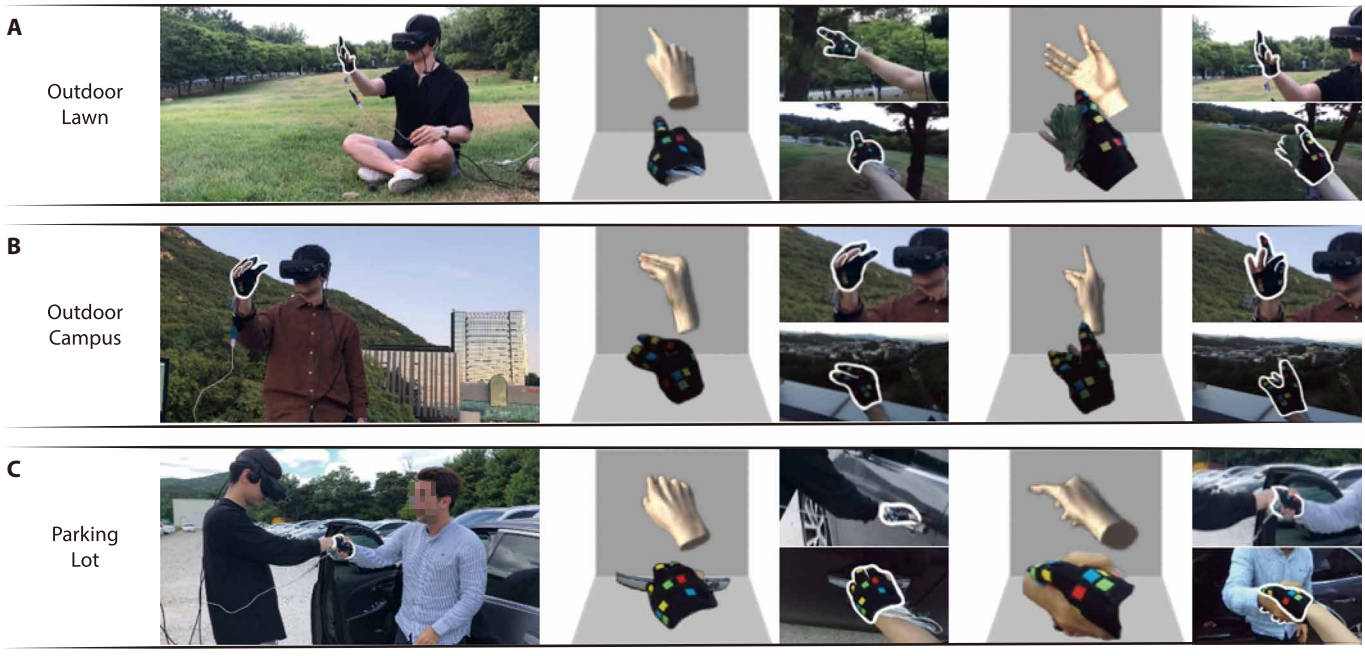



**Fig. 6. Qualitative evaluations in outdoor environments.** (**A**) Lawn. (**B**) Campus. (**C**) Parking lot.

trackers (*25*) for the wrist sensing. We thus conducted experiments for some such challenging outdoor scenarios: lawn, campus, and parking lot. For the campus/lawn scenarios, backgrounds and lighting conditions markedly differed from those of the indoors, making proper running of vision-based algorithms challenging. In the parking lot scenario, some everyday activities (opening car door and shaking hands) were also included, which rendered the hand tracking even more challenging due to the ferromagnetic materials of the vehicle (against IMU/compass tracking), mechanical contacts with people/objects (against soft sensor tracking), or severe occlusions


**Fig. 7. Qualitative evaluations for various occlusions.** (**A**) Outside the FOV. (**B**) Self-occlusion. (**C**) Severe occlusion from surroundings. (**D**) Interaction with various objects. (**E** and **F**) Robustness of VIST against magnetic interference/contact/occlusion from the tablet and comparison with other tracking methods (movie S7) (*9*, *22*).

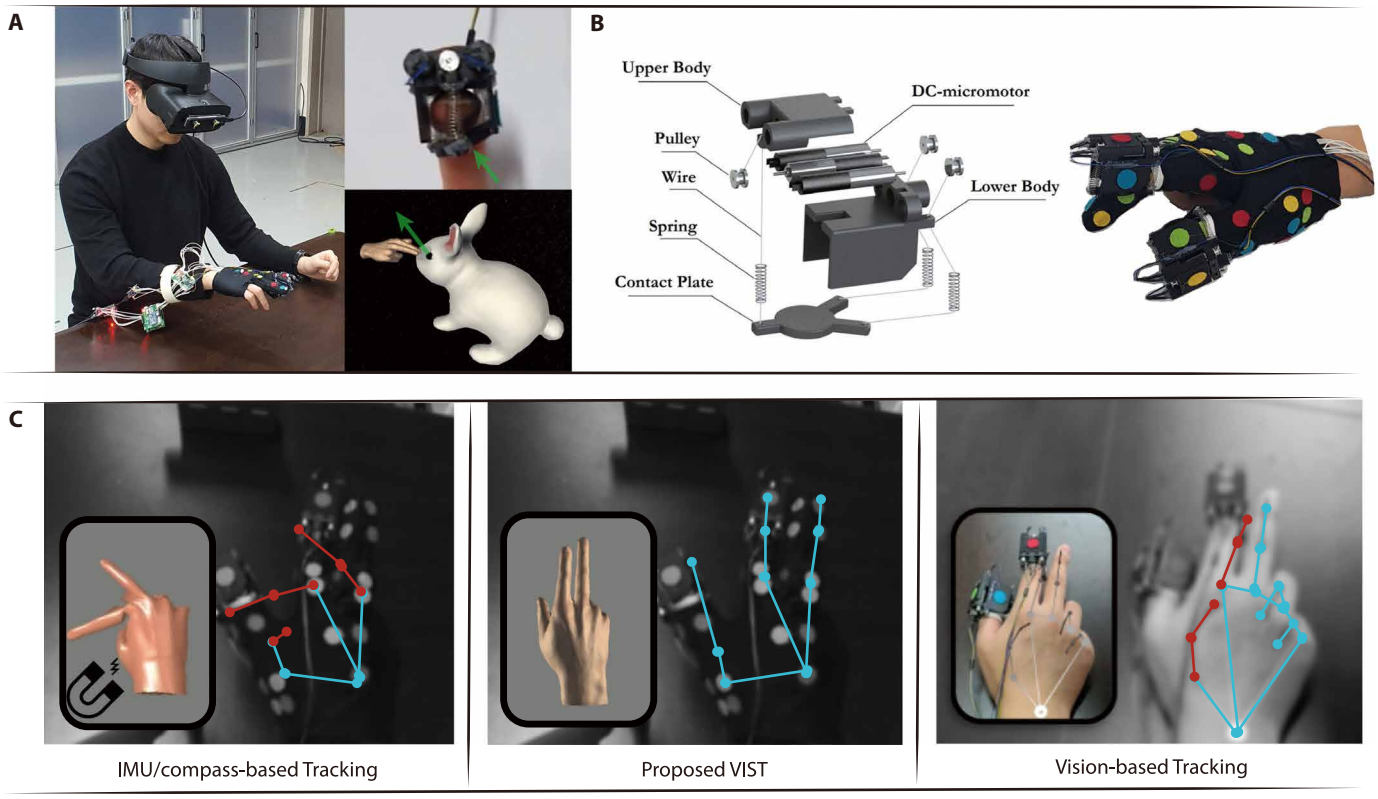



**Fig. 8. Qualitative experiments with CHDs in VR.** (**A**) Experimental setup with virtual rabbit. (**B**) 3-DOF CHDs and wearing configuration. (**C**) Robustness of VIST against visual distortion, mechanical contacts, and electromagnetic interference from CHDs and comparison with other tracking methods (movie S8) (*9*, *22*).

(against vision-based tracking). Our VIST system, in contrast, can robustly track the hand motion in all these scenarios (see Fig. 6 and movie S6).

***Self-occlusion***

We also performed qualitative evaluations for the cases of self occlusions and outside the FOV. The issue of self-occlusion is the fundamental limitation of the vision-based systems that still remained unsolved (*12*, *17*, *48*), whereas, when the hand is even partially outside the FOV, the vision-based tracking cannot articulate invisible hand segments or even cannot recognize the hand altogether (*10*, *48*). In contrast, our VIST can still accurately track self-occluded hand poses (e.g., fingers behind palm or middle finger behind index finger) (see Fig. 7B and movie S2). This reaffirms the robustness of our VIST framework against self-occlusions in accordance with the quantitative evaluation in Fig. 3E. Our VIST system can also track the invisible segments even outside the FOV (Fig. 7A) by using the IMU information with real-time autocalibrated hand anatomical parameters. This property is desirable for real applications, because users can freely move their hands without continuously paying attention to keep their hands within the camera FOV (e.g., for our system, the operable area increases by about 50% at a distance of 25 cm from the camera).

***Object interaction***

When human hands interact with objects, the issue of occlusions naturally arises (*13*, *17*, *48*, *70*). It is infeasible to include all everyday objects with accurate annotations in the training set. Even for objects in the training set, many systems still fail to track the hand motion if their size is large to induce occlusions (*17*, *13*). Such object interaction is also challenging for soft wearable hand tracking (*26*), because the soft sensors generally cannot distinguish the deformation signal of the hand motion from that of the object interaction. The IMU/compass-based wearable tracking systems (*22*) can also be compromised when interacting with objects containing ferromagnetic materials (e.g., metallic products and components with magnets) or internal electrical currents (e.g., powered tools and workstations), which can severely breach the compass signals.

In contrast, as shown in Fig. 7 (C and D) and movie S2, our VIST maintains accurate hand tracking even when the hand is occluded by, or interacting with, various objects/surroundings (e.g., under the table, behind ping-pong racket, and pressing gamepad buttons). We also conducted experiments to manipulate a tablet PC, which has embedded magnets and ferromagnetic materials (against IMU/compass tracking), a form factor prone to cause occlusions (against vision-based tracking), and contacts on many parts of the hand (against soft sensor tracking). Even with this tablet, our VIST system can retain the hand tracking (Fig. 7E), whereas vision-based and IMU/compass-based approaches fail (Fig. 7F). See also movie S7.

***Wearing fingertip CHD***

Vision-based hand tracking algorithms, which use datasets based on bare hands for the training, generally cannot track the hands well when the user wears devices/attachments on the hand. Soft wearable tracking is also vulnerable to those extra devices/attachments, because the soft sensor signals can be distorted by their contacts. Excessive deformation due to the device/attachment can even induce permanent signal offset in the soft sensors. Magnets or electrical actuators embedded in the device can severely interfere with

IMU/compass tracking as well. To verify the robustness of our VIST framework against such extra devices/attachments, we used the same CHDs (*22*) as used above (see also Fig. 8B). We then conducted a VR haptic exploration task, where human users can receive 3-DOF haptic feedback when touching the surface of a virtual rabbit (Fig. 8A). As shown in Fig. 8C and movie S8, vision-based and IMU/compass-based tracking systems become unstable with the CHDs due to the aforementioned reasons, whereas our proposed VIST can maintain stable and accurate hand tracking during the VR experiments.

## DISCUSSION

Through quantitative and qualitative evaluations, we demonstrate that our VIST framework operates robustly and with high performance in challenging real-world scenarios. In particuar, VIST enables the interaction of diverse objects with hand size/shape variability. In contrast, vision-based systems are not robust for untrained objects/hands (*19*) or object occlusions (*16*, *48*), soft sensor wearable systems are susceptible to object mechanical contacts (*26*, *28*, *29*), and IMU/compass or magnetic wearable systems (*20*, *22*, *23*, *32*, *35*) are fragile to ferromagnetic objects or electrical currents. Both the soft and IMU/compass wearable systems are also well known for their tip tracking error proportional to the skeleton size.

Human hands can interact with a myriad of objects in daily life with different hand configurations. The robust hand tracking of our VIST framework may lead to its broader applicability for wide varieties of real-world applications that have so for eluded existing approaches (e.g., daily monitoring for rehab and tool operation skill assessment). The accurate tracking of VIST with CHDs (*22*) could lead to applications that require extra wearable devices/attachments (e.g., VR/AR, telepicking with CHDs, and soft prosthesis) (*45*). We also verify that the VIST system can robustly track hand motion outdoors, which is tough for most existing systems, because sunlight interferes with many types of IR sensors [e.g., RGB-D camera (*48*, *67*) and external IR tracker (*25*) required for wearable tracking systems (*21*–*23*)], whereas outdoor hand tracking datasets for machine learning are fairly scarce. Our outdoor experiments verify not only the complete portability of the VIST system in terms of hardware/algorithm but also its feasibility for promising outdoor applications (e.g., intuitive interface for 3D drone swarm control).

The key reason of the superior performance of our VIST framework is that we alleviate the inherent issues of each sensor by visual-inertial TC fusion. The VIST framework circumvents the fundamental issues of vision-based systems [occlusion, generalization, and slow update (*19*, *17*, *48*)], because the motion of the occluded parts can still be accurately estimated by using the IMU information at a high rate (about 100 Hz) with the real-time/autocalibrated hand/sensor-related parameters, anatomical constraints, and still visible markers. Our VIST system also overcomes the issues of drift or magnetic interference of IMU/compass-wearable systems by exploiting the visual information in conjunction with the anatomical constraints and also the issues of unmodeled contacts for soft sensor wearable systems, because the camera and IMUs are immune to them. Moreover, the integrated autocalibration endows our VIST framework with improved accuracy and convenience as compared to existing IMU/compass or soft sensor wearable systems, where those parameters are calibrated once before the operation while the user takes several indicated poses, which is inevitable with human errors (*52*–*54*, *71*).

In conclusion, our VIST framework solves those fundamental limitations of existing hand tracking systems. This improved performance can be achieved by fusing the complementary aspects of visual and inertial sensors in TC fusion, which turns out to be crucial to properly address the peculiarity of the hand (and finger) tracking. With the ruggedness, portability, and affordable cost, our VIST system could allow for many promising real-world applications based on hand motion tracking.

## MATERIALS AND METHODS

### Sensor glove and stereo camera

We fabricated the sensor glove based on our previous work (*22*), whose sensor configuration was slightly modified such that seven IMUs are attached. We deployed low-cost commercial IMUs, MPU9250 (InvenSense), and connected them to a custom-built microcontroller unit (MCU) board based on ATmega328. This board collects the IMU data and sends it to the computer running the main algorithm. The data acquisition from each IMU is at 100 Hz, and the data are sent through the serial peripheral interface (SPI) communication protocol.

A user was recommended to wear a sensor glove slightly smaller than their hands to avoid sensor slippage on the hand. The glove itself was made with spandex fabric, which has sufficient elasticity to stretch with the human hand. Despite such elasticity, there is a limit on the hand size that a glove can cover, so we constructed two sensor gloves of different sizes (fig. S1). The length from the wrist to the middle finger tip and the width from the thumb MCP joint to the right side of hand are 18.3 and 9.4 cm for the smaller glove and 20.0 and 10.5 cm for the larger one. The total weights of the sensor gloves, including the markers, IMUs, and MCUs, are only 52 and 55 g, respectively.

We adopted color blobs (circular/square color patches made from fabric) as the visual passive markers because they can be simply attached to the gloves without extra electronic installations (power, wire, or diode). Other types of anonymous markers are equally possible for our proposed VIST framework, according to the operating environment (e.g., gloves made with pattern-printed fabrics, IR/ultraviolet light-emitting diodes, reflective markers with IR cameras, or deep learning–based features). The color blobs were attached to the designated positions of the sensor gloves (including positions directly above the IMUs), which were empirically determined to ensure that appropriate numbers of markers could be seen from any viewpoint of the camera. Fabric patch with four distinct hues (red, yellow, green, and blue) were used in fabricating the markers. We attached different shapes of markers (square and circle) to the two different gloves. The length of one side of square marker and diameter of the circular marker are both 12 mm, and a total of 37 color blobs were attached to each glove.

We used Stereolabs ZED Mini (*72*) as the stereo camera, which is manufactured with affordable weight, size, and baseline to be comfortably equipped with an HMD: 62.9 g weight, 90° (*H*) by 60° (*V*) FOV, 63-mm baseline, and 1280 pixel–by–720 pixel resolution for each image. Other types of vision sensors (e.g., monocular camera and depth/IR camera with IR markers) are also applicable for our proposed VIST framework by slightly modifying the marker detection/stereo matching processes. We assumed that the stereo camera is mounted on the user's head part (e.g., equipped with an HMD, AR glass, or safety goggle), because this configuration makes our system fully portable without installing external sensors and compatible

to commercial HMDs, which are normally equipped with two or more cameras (e.g., Oculus Quest and HoloLens).

### Experimental setup

In the quantitative evaluations, we used OptiTrack MOCAP system and attached IR reflective markers on the four keypoints (Fig. 3A and movie S10). Participants were instructed to sit on a chair surrounded by the MOCAP cameras, with the stereo camera positioned in front of the participant on a table facing downward at about 60°. The camera was fixed during the experiment so that the pure tracking errors of the hand motions relative to the camera {*C*} could be measured. This aligned with most existing vision-based studies, where they tracked the hand motion also with respect to the camera (*10*–*12*, *62*, *73*).

For the free motion quantitative evaluation, each participant was instructed to follow a hand image displayed on the monitor, which was randomly selected from the large and expansive (tens of thousands) compounded image datasets of (*10*, *61*). For the quantitative evaluation with object interaction, we used our own-built image datasets consisting of 100 images for each of the eight objects, whereas, for the quantitative evaluation with the haptic device, a virtual sphere with random location and size was displayed and the participant was instructed to touch it using their thumb or index finger with penetration-dependent normal haptic feedback. The time interval for the next random image was decided to be 3 s after performing a pilot test: If the time interval is too short, participants cannot follow the displayed images correctly; if it is too long, the tracking error is underestimated. Fifteen participants were recruited, all right-handed males in the age range of 22 to 31 years old with no known perception/movement disorders and various hand shapes (fig. S6). All the experiments were conducted in accordance with the Helsinki Declaration.

## SUPPLEMENTARY MATERIALS

www.science.org/doi/10.1126/scirobotics.abe1315
Notes S1 to S7
Figs. S1 to S10
Movies S1 to S10

## REFERENCES AND NOTES


1. Video Friday: Amazon CEO Jeff Bezos Tries Dexterous Robot Hands (2019); https://spectrum.ieee.org/automaton/robotics/robotics-hardware/video-friday-amazon-ceo-jeff-bezos-dexterous-robot-hands/.
2. J. Bimbo, C. Pacchierotti, M. Aggravi, N. Tsagarakis, D. Prattichizzo, Teleoperation in cluttered environments using wearable haptic feedback, in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems* (IEEE, 2017), pp. 3401–3408.
3. A. Tobergte, P. Helmer, U. Hagn, P. Rouiller, S. Thielmann, S. Grange, A. Albu-Schäffer, F. Conti, G. Hirzinger, The sigma. 7 haptic interface for mirosurge: A new bi-manual surgical console, in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems* (IEEE, 2011), pp. 3023–3030.
4. KUKA smartPAD teach pendant (2018); https://www.kuka.com/en-de/products/robot-systems/robot-controllers/smartpad/.
5. E. Ackernam, Your finger on a tablet can control entire swarms of robots (2015); https://spectrum.ieee.org/automaton/robotics/robotics-software/georgia-tech-robot-swarm-control/.
6. Oculus Quest 2 & Touch; www.oculus.com/quest-2/features/.
7. VIVE controller; www.vive.com/us/accessory/controller/.
8. Leap Motion; www.ultraleap.com/product/leap-motion-controller/.
9. F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C.-L. Chang, M. Grundmann, Mediapipe hands: On-device real-time hand tracking, in *Proceedings of IEEE International Conference on Computer Vision (Workshops)* (IEEE, 2019).
10. F. Mueller, F. Bernard, O. Sotnychenko, D. Mehta, S. Sridhar, D. Casas, C. Theobalt, Ganerated hands for real-time 3d hand tracking from monocular rgb, in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2018), pp. 49–59.
11. A. Spurr, U. Iqbal, P. Molchanov, O. Hilliges, J. Kautz, Weakly supervised 3d hand pose estimation via biomechanical constraints, in *Proceedings of European Conference on Computer Vision* (Springer, 2020), pp. 211–228.
12. G. Moon, Y. Ju, K. Lee, V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map, in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2018), pp. 5079–5088.
13. C. Zimmermann, T. Brox, Learning to estimate 3d hand pose from single rgb images, in *Proceedings of IEEE International Conference on Computer Vision* (IEEE, 2017), pp. 4903–4911.
14. C. Qian, X. Sun, Y. Wei, X. Tang, J. Sun, Realtime and robust hand tracking from depth, in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2014), pp. 1106–1113.
15. T. Sharp, C. Keskin, D. Robertson, J. Taylor, J. Shotton, D. Kim, C. Rhemann, I. Leichter, A. Vinnikov, Y. Wei, D. Freedman, P. Kohli, E. Krupka, A. Fitzgibbon, S. Izadi, Accurate, robust, and flexible real-time hand tracking, in *Proceedings of ACM Conference on Human Factors in Computing Systems* (Association for Computing Machinery, 2015), pp. 3633–3642.
16. S. Hampali, M. Rad, M. Oberweger, V. Lepetit, Honnotate: A method for 3d annotation of hand and object poses, *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2020), pp. 3196–3206.
17. S. Sridhar, F. Mueller, M. Zollhöfer, D. Casas, A. Oulasvirta, C. Theobalt, Real-time joint tracking of a hand manipulating an object from rgb-d input, in *Proceedings of European Conference on Computer Vision* (Springer, 2016), pp. 294–310.
18. S. Yuan, Q. Ye, B. Stenger, S. Jain, T.-K. Kim, Bighand2. 2m benchmark: Hand pose dataset and state of the art analysis, in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2017), pp. 4866–4874.
19. A. Armagan, G. Garcia-Hernando, S. Baek, S. Hampali, M. Rad, Z. Zhang, S. Xie, M. Chen, B. Zhang, F. Xiong, Y. Xiao, Z. Cao, J. Yuan, P. Ren, W. Huang, H. Sun, M. Hrúz, J. Kanis, Z. Krňoul, Q. Wan, S. Li, L. Yang, D. Lee, A. Yao, W. Zhou, S. Mei, Y. Liu, A. Spurr, U. Iqbal, P. Molchanov, P. Weinzaepfel, R. Brégier, G. Rogez, V. Lepetit, T.-K. Kim, Measuring generalisation to unseen viewpoints, articulations, shapes and objects for 3D hand pose estimation under hand-object interaction, in *Proceedings of European Conference on Computer Vision* (Springer, 2020), pp. 85–101.
20. Perception Neuron; https://neuronmocap.com/.
21. Manus VR; https://manus-vr.com/.
22. Y. Lee, M. Kim, Y. Lee, J. Kwon, Y. Park, D. J. Lee, Wearable finger tracking and cutaneous haptic interface with soft sensors for multi-fingered virtual manipulation. *IEEE/ASME Trans. Mech.* **24**, 67–77 (2019).
23. T. L. Baldi, S. Scheggi, L. Meli, M. Mohammadi, D. Prattichizzo, GESTO: A glove for enhanced sensing and touching based on inertial and magnetic sensors for hand tracking and cutaneous feedback. *IEEE Trans. Human Mach. Syst.* **47**, 1066–1076 (2017).
24. G. Santaera, E. Luberto, A. Serio, M. Gabiccini, A. Bicchi, Low-cost, fast and accurate reconstruction of robotic and human postures via imu measurements, in *Proceedings of IEEE International Conference on Robotics and Automation* (IEEE, 2015), pp. 2728–2735.
25. VIVE tracker; www.vive.com/us/accessory/controller.
26. Cyber Glove Systems; www.cyberglovesystems.com.
27. O. Glauser, S. Wu, D. Panozzo, O. Hilliges, O. Sorkine-Hornung, Interactive hand pose estimation using a stretch-sensing soft glove. *ACM Trans. Graphics* **38**, 1–15 (2019).
28. W. Park, K. Ro, S. Kim, J. Bae, A soft sensor-based three-dimensional (3-d) finger motion measurement system. *Sensors* **17**, 420 (2017).
29. D. H. Kim, S. W. Lee, H.-S. Park, Improving kinematic accuracy of soft wearable data gloves by optimizing sensor locations. *Sensors* **16**, 766 (2016).
30. T. K. Chan, Y. K. Yu, H. C. Kam, K. H. Wong, Robust hand gesture input using computer vision, inertial measurement unit (imu) and flex sensors, in *Proceedings of IEEE International Conference on Mechatronics, Robotics and Automation* (IEEE, 2018), pp. 95–99.
31. S. Han, T. Kim, D. Kim, Y.-L. Park, S. Jo, Use of deep learning for characterization of microfluidic soft sensors. *IEEE Robot. Autom. Lett.* **3**, 873–880 (2018).
32. HaptX Gloves; https://haptx.com/.
33. Polhemus; https://polhemus.com/motion-tracking/hand-and-finger-trackers/.
34. F. S. Parizi, E. Whitmire, S. Patel, Auraring: Precise electromagnetic finger tracking, in *Proceedings of ACM Conference on Interactive, Mobile, Wearable and Ubiquitous Technologies* (ACM, 2019), vol. 3, pp. 1–28.
35. Y. Ma, Z.-H. Mao, W. Jia, C. Li, J. Yang, M. Sun, Magnetic hand tracking for human-computer interface. *IEEE Trans. Magnet.* **47**, 970–973 (2011).
36. Dexmo; www.dextarobotics.com/en-us.
37. Sense Glove; www.senseglove.com.
38. P. Ben-Tzvi, J. Danoff, Z. Ma, The design evolution of a sensing and force-feedback exoskeleton robotic glove for hand rehabilitation application. *J. Mech. Robot.* **8**, 051019 (2016).
39. I. Sarakoglou, A. Brygo, D. Mazzanti, N. G. Hernandez, D. G. Caldwell, N. G. Tsagarakis, Hexotrac: A highly under-actuated hand exoskeleton for finger tracking and force feedback, *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems* (IEEE, 2016), pp. 1033–1040.

40. Y. Tao, H. Hu, H. Zhou, Integration of vision and inertial sensors for 3D arm motion tracking in home-based rehabilitation. *Int. J. Robot. Res.* **26**, 607–624 (2007).
41. G. Bleser, G. Hendeby, M. Miezal, Using egocentric vision to achieve robust inertial body tracking under magnetic disturbances, in *Proceedings of IEEE International Symposium on Mixed and Augmented Reality* (IEEE, 2011), pp. 103–109.
42. R. Mallat, V. Bonnet, M. A. Khalil, S. Mohammed, Upper limbs kinematics estimation using affordable visual-inertial sensors. *IEEE Trans. Autom. Sci. Eng.* **2020**, 1–11 (2020).
43. Y. Li, D. Weng, D. Li, Y. Wang, A low-cost drift-free optical-inertial hybrid motion capture system for high-precision human pose detection, in *Proceedings of IEEE International Symposium on Mixed and Augmented Reality (Adjunct)* (IEEE, 2019), pp. 75–80.
44. M. Andrychowicz, B. Baker, M. Chociej, R. Józefowicz, B. McGrew, J. Pachocki, A. Petron, M. Plappert, G. Powell, A. Ray, J. Schneider, S. Sidor, J. Tobin, P. Welinder, L. Weng, W. Zaremba, Learning dexterous in-hand manipulation. *Int. J. Robot. Res.* **39**, 3–20 (2020).
45. B. B. Kang, H. Choi, H. Lee, K.-J. Cho, Exo-glove poly II: A polymer-based soft wearable robot for the hand with a tendon-driven actuation system. *Soft Robot.* **6**, 214–227 (2019).
46. D. Kim, B. B. Kang, K. B. Kim, H. Choi, J. Ha, K.-J. Cho, S. Jo, Eyes are faster than hands: A soft wearable robot learns user intention from the egocentric view. *Sci. Robot.* **4**, eaav2949 (2019).
47. C. Wong, Z.-Q. Zhang, B. Lo, G.-Z. Yang, Wearable sensing for solid biomechanics: A review. *IEEE Sens. J.* **15**, 2747–2760 (2015).
48. F. Mueller, D. Mehta, O. Sotnychenko, S. Sridhar, D. Casas, C. Theobalt, Real-time hand tracking under occlusion from an egocentric rgb-d sensor. in *Proceedings of IEEE International Conference on Computer Vision* (IEEE, 2017), pp. 1284–1293.
49. A. Tkach, A. Tagliasacchi, E. Remelli, M. Pauly, A. Fitzgibbon, Online generative model personalization for hand tracking. *ACM Trans. Graph.* **36**, 1–11 (2017).
50. M. Santello, M. Flanders, J. F. Soechting, Postural hand synergies for tool use. *J. Neurosci.* **18**, 10105–10115 (1998).
51. C.-E. Hrabia, K. Wolf, M. Wilhelm, Whole hand modeling using 8 wearable sensors: Biomechanics for hand pose prediction, in *Proceedings of Augmented Human International Conference* (Association for Computing Machinery, 2013), pp. 21–28.
52. Q. Yuan, I.-M. Chen, A. W. Sin, Method to calibrate the skeleton model using orientation sensors, in *Proceedings of IEEE International Conference on Robotics and Automation* (IEEE, 2013), pp. 5297–5302.
53. H. J. Luinge, P. H. Veltink, C. T. Baten, Ambulatory measurement of arm orientation. *J. Biomech.* **40**, 78–85 (2007).
54. T. Seel, J. Raisch, T. Schauer, Imu-based joint angle measurement for gait analysis. *Sensors* **14**, 6891–6909 (2014).
55. A. Myronenko, X. Song, Point set registration: Coherent point drift. *IEEE Trans. Pattern Anal. Mach. Intell.* **32**, 2262–2275 (2010).
56. O. Hirose, A bayesian formulation of coherent point drift. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**, 2269–2286 (2021).
57. W. Gao, R. Tedrake, Filterreg: Robust and efficient probabilistic point-set registration using gaussian filter and twist parameterization, in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2019), pp. 11095–11104.
58. A. Aldoma, Z.-C. Marton, F. Tombari, W. Wohlkinger, C. Potthast, B. Zeisl, R. B. Rusu, S. Gedikli, M. Vincze, Tutorial: Point cloud library: Three-dimensional object recognition and 6 dof pose estimation. *IEEE Robot. Autom. Mag.* **19**, 80–91 (2012).
59. N. Trawny, S. I. Roumeliotis, "Indirect kalman filter for 3D attitude estimation" (Techical Report, University of Minnesota, 2005), vol. 2.
60. S. M. Weiss, "Vision based navigation for micro helicopters," thesis, ETH Zurich (2012).
61. J. Zhang, J. Jiao, M. Chen, L. Qu, X. Xu, Q. Yang, A hand pose tracking benchmark from stereo matching, in *Proceedings of IEEE International Conference on Image Processing* (IEEE, 2017), pp. 982–986.
62. Y. Yang, D. Ramanan, Articulated human detection with flexible mixtures of parts. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 2878–2890 (2012).
63. Y. Lee, I. Jang, D. J. Lee, Enlarging just noticeable differences of visual-proprioceptive conflict in vr using haptic feedback, in *Proceedings of IEEE World Haptics Conference* (IEEE, 2015), pp. 19–24.
64. H. Z. Tan, M. A. Srinivasan, C. M. Reed, N. I. Durlach, Discrimination and identification of finger joint-angle position using active motion. *ACM Trans. Appl. Percept.* **4**, 10 (2007).
65. Z. Zhang, S. Xie, M. Chen, H. Zhu, Handaugment: A simple data augmentation method for depth-based 3D hand pose estimation (2020); arXiv:2001.00702 [cs.CV] (3 January 2020).
66. W. Huang, P. Ren, J. Wang, Q. Qi, H. Sun, Awr: Adaptive weighting regression for 3D hand pose estimation. *Proc. AAAI Conf. Artificial Intell.* **34**, 11061–11068 (2020).
67. S. Sridhar, F. Mueller, A. Oulasvirta, C. Theobalt, Fast and robust hand tracking using detection-guided optimization, in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2015), pp. 3213–3221.
68. M. Meilland, A. I. Comport, P. Rives, Dense omnidirectional rgb-d mapping of large-scale outdoor environments for real-time localization and autonomous navigation. *J. Field Robot.* **32**, 474–503 (2015).
69. S. M. Abbas, A. Muhammad, Outdoor RGB-D SLAM performance in slow mine detection, in *Proceedings of German Conference on Robotics* (VDE, 2012), pp. 1–6.
70. P. Panteleris, A. Argyros, Back to rgb: 3d tracking of hands and hand-object interactions based on short-baseline stereo, in *Proceedings of IEEE International Conference on Computer Vision (Workshops)* (IEEE, 2017).
71. Xsense; www.xsens.com.
72. Zed Mini; www.stereolabs.com/zed-mini.
73. U. Iqbal, P. Molchanov, T. B. J. Gall, J. Kautz, Hand pose estimation via latent 2.5D heatmap regression, in *Proceedings of European Conference on Computer Vision* (Springer, 2018), pp. 118–134.


**Acknowledgments:** We thank S.-M. Lee for assistance with performing the quantitative experiments, Y.-S. Lee for assistance with building VR environments, J. J.-R. Song for working on the graphical presentations of the figures, D.-G. Min for assistance with implementing the collaborative robotic arm interaction, Y.-H. Lee for assistance with fabricating the sensor glove, and J.-Y. Yoon for assistance with performing the experiments. **Funding:** This research was supported by two National Research Foundation of Korea (NRF) grants funded by the Korean government (MSIT) (nos. NRF-2016R1A5A1938472 and NRF-2020R1A2C3010039). **Author contributions:** Y.L. designed and built the hardware and software, performed and analyzed the experiments, and wrote the manuscript. W.D. assisted with building the hardware and software as well as writing the manuscript. H.Y. assisted with performing the experiments and writing the manuscript. J.H. assisted with the fabrication of the haptic devices and the implementation of the swarm drone interface. W.L. assisted with the design and fabrication of the sensor glove. D.L. directed the project and edited the manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to support the conclusions of this manuscript are included in the main text or Supplementary Materials and have been deposited in the database https://doi.org/10.7910/DVN/T7ASAC.

Submitted 15 October 2020
Accepted 8 September 2021
Published 29 September 2021
10.1126/scirobotics.abe1315

**Citation:** Y. Lee, W. Do, H. Yoon, J. Heo, W. Lee, D. Lee, Visual-inertial hand motion tracking with robustness against occlusion, interference, and contact. *Sci. Robot.* **6**, eabe1315 (2021).

Science Robotics

# Visual-inertial hand motion tracking with robustness against occlusion, interference, and contact

Yongseok Lee, Wonkyung Do, Hanbyeol Yoon, Jinuk Heo, WonHa Lee, and Dongjun Lee

*Sci. Robot.* **6** (58), eabe1315. DOI: 10.1126/scirobotics.abe1315







*Science Robotics* (ISSN 2470-9476) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science Robotics* is a registered trademark of AAAS.