공학박사학위논문

인간 기계 상호작용을 위한 강건하고 정확한 손동작 추적 기술 연구

Robust and Accurate Hand Tracking for Real-World Human Machine Interaction

2021 년 8 월

서울대학교 대학원

기계항공공학부

이 용 석

인간 기계 상호작용을 위한 강건하고 정확한 손동작 추적 기술 연구

Robust and Accurate Hand Tracking for Real-World Human Machine Interaction

지도교수 이 동 준

이 논문을 공학박사 학위논문으로 제출함

2021년 4월

서울대학교 대학원

기계항공공학부

이 용 석

이용석의 공학박사 학위논문을 인준함

2021년 6월

위 원 장 : 박 종 우

부위원장 : 이 동 준

위 원 : 조 규 진

위 원 : 한 보 형

위 원 : 배 준 범

[redacted]

Robust and Accurate Hand Tracking for Real-World Human Machine Interaction

Yongseok Lee
Department of Mechanical & Aerospace Engineering
Seoul National University

Abstract

Hand-based interface is promising for realizing intuitive, natural and accurate human machine interaction (HMI), as the human hand is main source of dexterity in our daily activities. For this, the thesis begins with the human perception study on the detection threshold of visuo-proprioceptive conflict (i.e., allowable tracking error) with or without cutantoues haptic feedback, and suggests tracking error specification for realistic and fluidic hand-based HMI. The thesis then proceeds to propose a novel wearable hand tracking module, which, to be compatible with the cutaneous haptic devices spewing magnetic noise, opportunistically employ heterogeneous sensors (IMU/compass module and soft sensor) reflecting the anatomical properties of human hand, which is suitable for specific application (i.e., finger-based interaction with finger-tip haptic devices). This hand tracking module however loses its tracking when interacting with, or being nearby, electrical machines or ferromagnetic materials. For this, the thesis presents its main contribution, a novel visual-inertial skeleton tracking (VIST) framework, that can provide accurate and robust hand (and finger) motion tracking even for many challenging real-world scenarios and environments, for which the state-of-the-art technologies are known to fail due to their respective fundamental limitations (e.g., severe occlusions for tracking purely with vision sensors; electromagnetic interference for tracking purely with IMUs (inertial measurement units) and compasses; and mechanical contacts for tracking purely with soft sensors). The proposed VIST framework comprises a sensor glove with multiple IMUs and passive visual markers as well as a head-mounted stereo camera; and a tightly-coupled filtering-based visual-inertial fusion algorithm to estimate the hand/finger motion and auto-calibrate hand/glove-related kinematic parameters simultaneously while taking into account the hand anatomical constraints. The VIST framework exhibits good tracking accuracy and robustness, affordable material cost, light hardware and software weights, and ruggedness/durability even to permit washing. Quantitative and qualitative experiments are also performed to validate the advantages and properties of our VIST framework, thereby, clearly demonstrating its potential for real-world applications.

Keywords: Human motion tracking, tightly-coupled sensor fusion, extended Kalman filtering, correspondence search, IMU, computer vision, human-computer interaction, human-robot interaction, virtual reality, augmented reality, haptic feedback

Student number: 2013-23082

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1. Motivation

Dexterous use of hands (with fingers) is the defining characteristics of human beings. Integrating the hand would then drastically improve efficiency, intuitiveness and richness of many real-world human machine (computer/robot) interaction (HMI) applications, including: 1) VR (virtual reality) and AR (augmented reality), where using the hand would provide substantially richer and real-life like experience as compared to the currently dominating 6-DOF "fist-based" interface or fingertip/gesture-based interfaces; and 2) robotic-hand haptic teleoperation (Fig. 1.1A), particularly that of anthropomorphic robotic hands ((Bimbo *et al.* 2017)), where a remote user can fully utilize their hand and fingers with some haptic feedback for complex manipulation tasks, instead of relying on (typically only up to 6-DOF (degree-of-freedom)) conventional haptic devices (Tobergte *et al.* 2011); 3) collaborative robot interaction (Fig. 1.1B), where a user can quickly and intuitively provide rich commands and cues to the robot using their hand and fingers, thereby, can

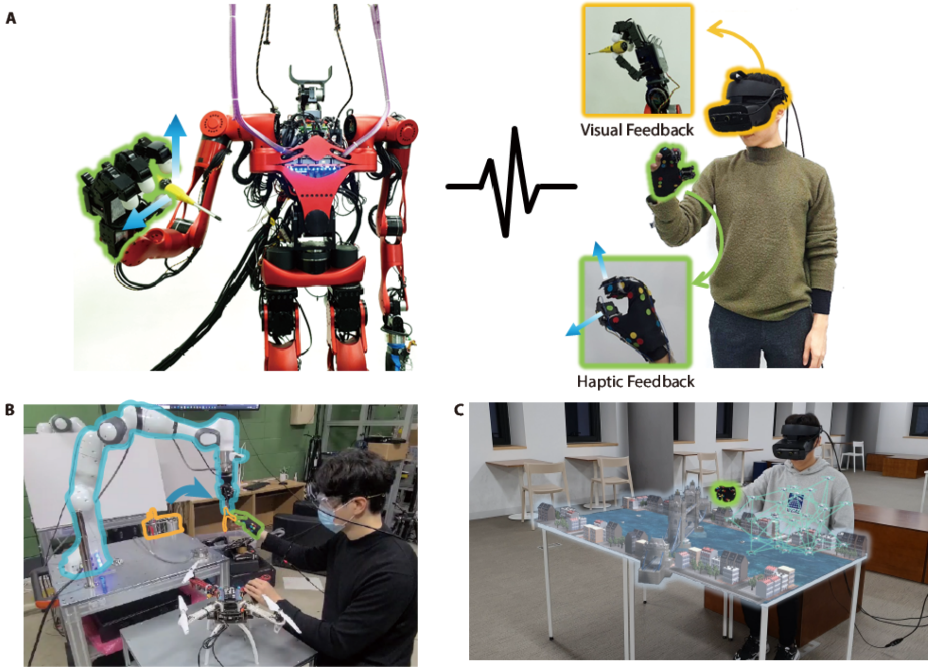


Figure 1.1. **(A)** Robotic hand teleoperation: a remote user can utilize their hand/fingers with haptic feedback for manipulation tasks of the humanoid hand (courtesy of DYROS, SNU). **(B)** Collaborative robot interaction: the robotic-arm assists in assembly task by delivering a necessary tool based on gesture recognition. **(C)** 3D swarm drone interface: users can efficiently control the complex formation of swarm or define virtual walls to dangerous regions.

make the interaction much safer and more fluidic as compared to the case of conventional pendant programming; and 4) 3D (three-dimensional) drone swarm control (Fig. 1.1C), where a field 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.

One of the key prerequisites to fully integrate this human hand into real-world engineering applications is to track the pose and configuration of the hands (with their fingers), accurately, robustly and affordably, which is still far from being materialized, since: 1) the human hand is of relatively a small size (i.e., requiring sub-centimeter level accuracy) with the five fingers, each exhibiting a complex and high-DOF (degree-of-freedom) motion by itself (e.g., thumb), so that the hand configuration cannot be completely determined only by a small number of perspective planes (e.g., cameras) while excluding their self-occlusion; 2) the human hand is supposed to interact with diverse objects, directly or indirectly related to the application (e.g., receiving a cell phone during gaming, tools for smart factory) in an environment, that may be dynamic (e.g., ambient lighting during day and night) and cluttered with possibly adversarial objects (e.g., multitude of colorful objects, tablets or machines interfering magnetic field); and 3) the human hand is all different (i.e., different size, shape, color, mechanical property, etc.) and how to use it even for the same application is also in general all different among users, making it very difficult to robustly anticipate its motion via interpolation or learning.

Numerous results have been proposed for this problem, and they may be categorized into the following three approaches, each, yet, with their respective fundamental limitations; occlusion (vision-based tracking), magnetic-interference (IMU-based tracking), and mechanical contact (soft wearable tracking), which will be explained more in the next section. Therefore, the main purpose of this thesis is to develop a novel hand tracking system, which overcomes every issue of existing systems and provides accurate and robust tracking results suitable for every real-world application. Before the development of the novel hand tracking system, the allowable error of hand tracking system is clarified in this thesis via human subject studies, which would be a crucial design specification for every hand tracking system. In addition, as shown in many studies about hand tracking, realistic cutaneous haptic feedback is desirable for finger-based interaction in terms of user experience (realness or preference), fatigue, or task performance (accuracy and completion time). The problems of current hand tracking methods for integration with the cutaneous haptic devices are that the devices normally raise issues of visual distortion (vision-based systems), magnetic-interference from embedded magnets and operating current (IMU-based system), or contact from the device (soft wearable system), which leads to the unstable tracking performance of existing methods. Thus a novel wearable hand tracking framework compatible with the wearable cutaneous haptic devices is developed in this thesis utilizing soft and IMU/compass sensors opportunistically. Expending the above hand tracking result, which is limited for the usage of haptic devices, thus, lose its generality for every fundamental issue of real-world scenarios, we propose a novel visual-inertial sensor fusion framework, which is accurate, robust, and affordable, thus, suitable for many real-world engineering applications in daily life.

1.2. Related Work

- **Visual-proprioceptive conflict of hand tracking**

  The imperfection of proprioceptive sense of humans is revealed in (Van Beers *et al.* 1998), which implies that the functioning of a finger-tracking system would still be proper if its tracking error can be made below a certain detection threshold of visual-proprioceptive conflict of users (Welch & Foxlin 2002). Several results have been proposed on the visual-proprioceptive conflict: static orientation error (Madsen & Stenholt 2014), effect of latency and noise (Livingston & Ai 2008), and drift angle of the arm (Burns *et al.* 2005) in virtual or mixed reality. It was shown in (Holmes & Spence 2005) that the position perception of the human is determined by a weighted sum of visual, proprioceptive, and other senses, implying that the haptic feedback would be able to affect the perception of finger-tracking error as also aimed for in this thesis. This interplays between haptic and other senses are also studied: the role of haptic and visual senses in curvature perception (Drewing & Ernst 2006), pseudo-haptic effects by modifying visual cues (Jang & Lee 2014), and the effect of matching visual cues with haptic cues on modifying felt position of subjects (Folegatti *et al.* 2009). Yet, to our knowledge, quantitative (e.g., detection threshold) analysis of the visual-proprioceptive conflict for complex 3D spatial motion and, further, quantitative analysis of the effect of haptic feedback on the visual-proprioceptive conflict threshold has not been explored before.

- **Vision-based hand tracking**

Vision-based hand tracking typically utilize a RGB (red-green-blue), stereo, or RGB-D (depth) camera to estimate the hand pose and configuration (e.g., (Zhang *et al.* 2019*a*; Mueller *et al.* 2018; Iqbal *et al.* 2018; Moon *et al.* 2018; Zimmermann & Brox 2017)). They reconstruct the configuration of hand even in a marker-less fashion while exploiting the recently-burgeoning machine learning techniques (e.g., convolution neural network (CNN)), thus, also show the improved tracking results (Zimmermann & Brox 2017; Iqbal *et al.* 2018; Moon *et al.* 2018) than classical vision-based hand tracking (Oikonomidis *et al.* 2011; Tompson *et al.* 2014). In addition, the study (Moon *et al.* 2018) design a novel 3D network structure for the voxelized data from the RGB-D sensor data, which provides improved tracking results than other RGB-D-based systems. To address the generalization issues of machine learning, the generative adversarial network (GAN) is used recently for hand tracking (Mueller *et al.* 2018), which augment the training dataset utilizing an artificial hand avatar. For more generality, creations of a vast data set of hand images, which is difficult due to the accurate annotations of many joints, are also studied using multi-cameras or electro-magnetic trackers (Hampali *et al.* 2020; Sridhar *et al.* 2016; Yuan *et al.* 2017). However, vision-based hand tracking still cannot circumvent the fundamental issue of occlusion (e.g., error drastically increased with some parts of hands in self-occlusion or outside camera FOV (field-of-view)) (Moon *et al.* 2018; Mueller *et al.* 2017). Moreover, the machine learning techniques are well-known for the issues of generalization (e.g., large error or even diverging possible for hand size/posture with objects outside the training set), which is also addressed in the recent hand tracking challenge (Armagan *et al.* 2020).

- **Soft wearable hand tracking**

  Soft wearable hand tracking employ a multitude of soft sensors, each producing signal according to their deformation, which are wrapped around the hand to estimate the hand and finger configuration (e.g., (Glauser *et al.* 2019; Park *et al.* 2017; Kim *et al.* 2016; Muth *et al.* 2014; Chossat *et al.* 2015)). Flexion/extension of the finger joints can be measured via stretchable soft sensors embedded on them (Chossat *et al.* 2015; Muth *et al.* 2014). The multi-DOFs joints (e.g., abduction/adduction of carpometacarpal (CMC) or metacarpophalangeal (MCP)) are also estimated by fabricating sophisticatedly sensor structure using conductive liquid metal materials. Optimal placement method of soft sensors is suggested (Kim *et al.* 2016) to estimate CMC joint and a deep network which utilizes the spatial structure of the sensor placement and per-user calibration method is proposed in (Glauser *et al.* 2019). However, all soft wearable hand tracking, which only measures the relative motions of adjacent joints, require with some extra exocentric sensor (e.g., camera) to measure their pose in the 3D space (Glauser *et al.* 2019). They also suffers from its inability to distinguish the motion-induced deformation from that induced by the contact, making it unsuitable for such applications as smart factory (with tool-holding) and prosthesis (with daily-life objects), which leads to the difficulty of calibration (due to the complex coupling among the soft sensors) and the limited ruggedness (e.g., permanent offset from large bending, squashing, or washing) (Lee *et al.* 2019).

- **IMU[1]-based wearable hand tracking**

  IMU-based hand tracking typically utilize IMU/compass modules, which are attached to every target link of the hand (e.g., (Lee *et al.* 2019; Baldi *et al.* 2017; Santaera *et al.* 2015; Kortier *et al.* 2014; Mizera *et al.* 2019)), to measure their 3-DOF (absolute) orientation. The nonlinear complementary filter (Mahony *et al.* 2008; Hrabia *et al.* 2013) or extended kalman filter (Kortier *et al.* 2014) are mostly utilized for estimating each segment rotation. Erroneous disturbances such as linear acceleration or magnetic distortion is compensated by the proposed filter (Lee *et al.* 2012; Roetenberg *et al.* 2005). To track the angles of 1-DOF joints (distal/intermediate joints), soft sensors are together utilized, which can estimate 1-DOF motion in simple structure (Lee *et al.* 2019; Mizera *et al.* 2019). This wearable tracking system is applied to real-world applications such as haptic interaction (Baldi *et al.* 2017; Lee *et al.* 2019). However, due to the drift of IMU, the wrist position cannot be estimated, which requires some additional sensor added to refurnish its pose information. However, IMU-based hand tracking is particularly susceptible to a change or interference of magnetic field, as it directly baffles the magnetometer (Roetenberg *et al.* 2005; Lee *et al.* 2019), thus, impossible to use when near or in contact with ferromagnetic objects or magnetism-spewing machines (e.g., steel wall, powered tool,

[1] IMU (inertial measurement unit) typically refers to a combination of a 3-axis accelerometer and a 3-axis gyroscope. In this thesis, we however use this term, IMU, to refer to a combination of a 6-axis IMU and a 3-axis magnetometer, since, for the hand tracking, this 9-axis IMU sensor configuration is almost always adopted.

cell phone, table, laptop).

- **Hand tracking compatible with fingertip haptic device**

  On top of the finger/hand tracking, haptic feedback is also imperative for immersive VR experiences. For this, cutaneous haptic feedback (with skin deformation) has received wide attention for wearable finger-based haptics for its portability, small form-factor and affordability as compared to kinesthetic haptic feedback (Kuchenbecker *et al.* 2008; Chinello *et al.* 2015, 2017), and its ability to supplement or even substitute the kinesthetic feedback (which is not suitable for wearable finger-based interaction) is studied (Prattichizzo *et al.* 2012*a*; Jang & Lee 2014; Quek *et al.* 2015). Various finger-tip cutaneous haptic devices have been proposed (e.g., (Minamizawa *et al.* 2007; Prattichizzo *et al.* 2010*a*; Meli *et al.* 2013; Leonardis *et al.* 2015; Schorr & Okamura 2017)). A two-DOF band-driven cutaneous device is proposed in (Minamizawa *et al.* 2007), and later adopted in (Prattichizzo *et al.* 2010*a*) and extended to a three-DOF wire-driven module (Meli *et al.* 2013). The work of (Meli *et al.* 2013) attach three FSR sensors on a contact plate and use their readings. Combining hand tracking modules with these cutaneous haptic devices has been researched to build a novel intuitive/dexterous/realistic interface for human machine interaction. The IMU/compass modules are attached on each segment of fingers and 1-DOF voice coil motors are employed on the fingertips (Baldi *et al.* 2015, 2017), with utilizing the simple structure of voice coil motors. However, the motors, which leads to the magnetic-interference to the compass, simply exerts 1-DOF force, which cannot deliver shear force. The work of (Weber *et al.* 2016) or MANUS VR glove provides finger tracking and haptic feedback simultaneously with IMUs and soft sensor adopted. However, the device of (Weber *et al.* 2016) utilizes only a single IMU on the dorsum of the hand and a single soft sensor for each finger, thus, not able to fully track large-DOF complex finger motions, and deliver only single-DOF vibrotactile feedback on the finger-tip, too simple to capture most of real-life finger/hand interactions. Vision-based hand tracking systems are also used to estimate hand motions with cutaneous haptic devices, several attempts have been made to integrate the RGB-D sensors or fiducial markers to wearable finger-based haptic systems (Frati & Prattichizzo 2011; Scheggi *et al.* 2015; Maisto *et al.* 2017), which also fundamental issues of occlusions from haptic devices.

- **Vision-inertial sensor fusion for hand/skeleton tracking**

  The sensor fusion of vision and inertial sensor are widely studied in robotics, which can be classified into the two categories, accurate/robust but complex tightly-coupled (TC) fusion (Shen *et al.* 2013; Hesch *et al.* 2014) and simpler but less accurate/robust loosely-coupled (LC) fusion (Weiss 2012; Forster *et al.* 2016), and applied for mapping/localization of robots such as unmanned ground vehicles (UGV) (Zhang *et al.* 2019*b*; Chen *et al.* 2019) or drones (Lynen *et al.* 2013; Faessler *et al.* 2016). The LC-fusion approach (e.g., only IMU drift correction, or just add two separate information (Chan *et al.* 2018)). The study (Bleser *et al.* 2011) uses 2 IMUs for each arm with a chest camera-IMU module, which detect the wrists by a blob with different colors (red/green) to correct the drift of the IMU. It use vision to overcome the magnetic disturbance, yet, not directly applicable to the problem of hand/finger tracking with many anonymous markers. The study in (Chan *et al.* 2018), which is not really tracking but hand gesture recognition using RGB-D camera aided with the IMU information. The AR marker is also utilized for sensor fusion (Trindade *et al.* 2012), but just a rigid-body (i.e, palm) not skeletal tracking of hand, thus, not applicable to the hand/finger tracking, and (Zhou *et al.* 2013) also has similar a rigid-body tracking only for trajectory tracking, utilizing single camera rather than AR marker. Extending the area from human hand to upper limb, (Tao *et al.* 2007) utilize single skeleton tracking (upper limb) with multiple IMUs and a single marker at the wrist in a loosely-coupled manner, since correspondence search not that challenging only for a single visual marker. In the study (Mallat *et al.* 2020), three AR-marker/IMU modules are used to track the motion of human upper limb motion, where the correspondence search can be easily done by AR-marker, whose size is too large to be applicable to the case of hand/finger tracking (with very high-resolution camera, the tracking rate would be much slower). In terms of whole-body tracking, the visual-inertial sensor fusion is studied in (Li *et al.* 2019) where 6 VIVE (optical) trackers and 17 non-axis IMU sensors are used where the optical trackers are mainly used to correct the drift of the IMU sensors, VIVE trackers, yet, not applicable to the hand tracking, which requires many small anonymous trackers within the small human hand. In summary, all the approaches of skeletal tracking rely only on the LC approach (e.g., just add two separate information) rather than TC-fusion, since this LC-fusion suffices for them due to small number of tagged markers (e.g., AR markers (Mallat *et al.* 2020), VIVE trackers (Li *et al.* 2019), distinguishable color markers (Bleser *et al.* 2011; Tao *et al.* 2007)) are utilized. However, to robustly/accurately track all large-DOF motions of small-size hands without problematic compass, correspondence search of many anonymous marker is required, which necessitate TC approach for visual-inertial sensor fusion in hand tracking problem.

1.3. Contribution

We begins this thesis with the research about the detection (absolute) threshold (Colman 2015) of the visual-proprioceptive conflict when performing finger-based operation in the chapter 2 This detection threshold is important for developing hand tracking system, since human would perceive the tracking error mainly depending on this as mentioned in the section 1.2. We also study about the effect of cutaneous haptic feedback to the detection threshold. To our knowledge, quantitative analysis of the visual-proprioceptive conflict for hand tracking system has not been explored, thus, we firstly identify the quantitative detection threshold ($5.11cm$) and realistic haptic device can alleviate the human perception of tracking error ($6.05cm$) by human subject study in the chapter, which would be one of crucial design specifications for every hand tracking system which currently exists or will be developed in the future.

In the chapter 3, we then propose a wearable finger tracking module, which overcomes issues of existing hand tracking systems when integrating with wearable cutaneous haptic devices. Concept of integrating hand tracking module with these cutaneous haptic devices would be promising, yet, the problem is that the integration normally generate issues of visual distortion (vision-based systems) or magnetic-interference from embedded magnets and operating current (IMU/compass wearable systems), which cannot be stably tracking by mentioned existing methods. Thus, we develop a novel hand tracking module, which optimally employ the heterogeneous sensors (IMU/compass and soft sensors) with taking into account the anatomical properties of human hand and magnetic disturbed range from attachment of the cutaneous haptic devices. A novel wearable cutaneous haptic interface is constructed via successful integration, and the validity and efficacy of the proposed finger tracking module is verified through a real VR manipulation task (i.e., peg-in-hole task).

At last, improving the above hand tracking modules, which partially solves issues of hand tracking specifically in case of wearing haptic devices, we finally develop a novel visual-inertial hand tracking framework (Fig. 1.2), which would not lose its generality for every fundamental issues of real-world scenarios in the chapter 4. One of the key contributions of our framework is that the sensor fusion of visual and inertial sensors in a TC-manner (coexistence of visual-to-inertial (e.g., IMU drift correction) and inertial-to-visual (e.g., IMU-aided correspondence search in section 4.4) which can address the peculiarity of hand tracking, that is, a number of segments by utilizing numerous passive markers. To provide rich information of the high-DOFs hand motions without compass while packed in a small form-factor of hand, we attach a large number of anonymous/passive markers on a wearable sensor glove. Then, the correspondence search of many anonymous markers even with occlusion cases becomes is very challenging, thus we develop the TC-fusion based sensor fusion framework in this thesis. To our knowledge, our VIST framework

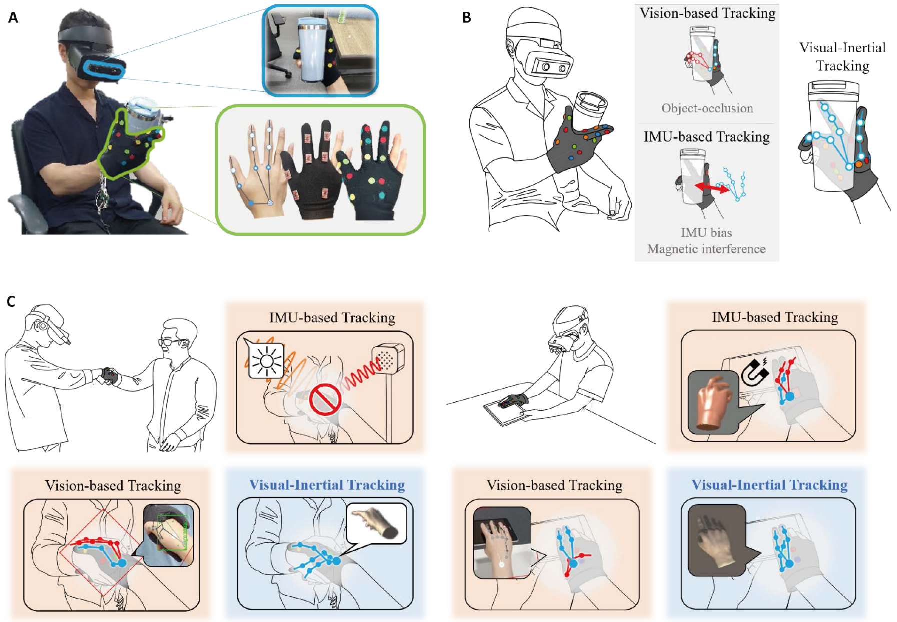


FIGURE 1.2. **(A)** Overview of hardware setup: the sensor glove (green) with two layers of the IMUs and visual markers and the stereo camera (blue). **(B)** Principle of our algorithm utilizing complementary aspects of the visual and inertial sensors. **(C)** Robustness of our system compared to conventional hand tracking systems (a vision-based tracking system (Zhang *et al.* 2019*a*) and a wearable tracking system (Lee *et al.* 2019)) under challenging scenarios: (Left) Handshaking with another person outdoors, which causes severe occlusion and sunlight interference to external IR sensor employed from many wearable tracking systems. (Right) Manipulating an electronic device, which causes severe occlusion and magnetic interference by embedded magnets or ferromagnetic materials.

is the very first result, that brings in the TC-fusion into the hand tracking, thereby, achieving the unprecedentedly accurate and robust tracking performance as reported in this thesis.

Some of important advantages and properties of our VIST framework, which will be described and verified in this thesis, can be summarized as follows:

- superior tracking accuracy due to the TC visual-inertial fusion as compared to other state-of-the-art approaches;
- robustness against occlusions, visually complex/changing environments and ambient lighting;
- robustness against electromagnetic interference, mechanical contacts, object manipulation and wearing devices;
- use convenience with the real-time/auto-calibration of anatomical/glove kinematic parameters integrated into the VIST algorithm; and
- ruggedness (e.g., washable by hand), affordability (e.g., total material cost $\approx$ \$100) and wearability (e.g., light weight (52–55 g)).

Our VIST framework may also be used to collect human data for the development of reinforcement learning (learning-by-demonstration) strategy for robotic object manipulation (Andrychowicz *et al.* 2020); or as a tracking module for the feedback-control of soft robotic hand prostheses (Kang *et al.* 2019; Kim *et al.* 2019). Our VIST framework can further be used for robotic systems with limbs, for which typical proprioceptive sensors are impossible to deploy (e.g., very thin tendon-driven robot with frequent impact, soft multi-legged robots with whole-body contact, etc.).

Chapter 2

Detection Threshold of Hand Tracking Error

2.1. Motivation

One of the key challenges for intuitive/dexterous finger-based interaction is how to track the pose of the fingers, which is in fact relevant to all the finger-based user interfaces in the area of general human-computer interaction (Pavlovic *et al.* 1997). Tremendous types of sensors and algorithms are employed and researched for hand/finger motion tracking. The problem is, regardless of which finger-tracking techniques are used, it is in general impossible to completely eliminate the tracking error due to the imperfection of sensor and algorithm. We describe the sources of tracking error of major currently-available finger tracking systems in the following paragraphs as more detailed description about them than the introduction.

Fundamentally, the vision-based methods suffer from the occlusion problem, i.e., the ray behind objects cannot reach the camera lens. Especially, since hand has dexterous finger motion with large Degree of freedoms (DOFs) in small space, a segment of hand is frequently occluded by others, which is referred to as self-occlusion. While tracking the self-occluded segment has been longstanding issues due to the absence of visual information for the occluded parts (Mueller *et al.* 2018; Zimmermann & Brox 2017; Sridhar *et al.* 2015), even state-of-the-art methods (Moon *et al.* 2018) show high tracking error on the occluded joints. Another occlusion problem would occur when some parts of hand is outside the field of view (FOV) of camera. The pose of the unobserved segments surely cannot be estimated, worse, simply recognizing human hand would become impossible since most of machine-learning-based methods assume entire silhouette of hand is observed. Given the fact that most of commercial cameras have fairly limited FOV, this outside the FOV issue inevitably restricts the range of hand motion, which can substantially degrade user experience of hand tracking system in practical usage. Moreover, vision-sensor has inferior performance in terms of sampling rate than IMU or soft sensors. Slower update-rate (normally up to 30 Hz) relative to hand motions also cause delayed estimation result or motion blur images. The resolution of vision sensor also cause the inevitable error for tracking algorithm.

Compared to this sensor, IMU-based wearable systems has higher frame rate (with some proper sensor fusion), immunity to the issue of occlusion (i.e., free from the line-of-sight requirement). The unmodeled signal/perturbation to acceleration/magnetometer measurement (e.g., linear acceleration and external magnetic distortion) substantially distort the hand tracking result as well. Moreover, the principle of IMU measuring the inertial properties of the attached rigid body (i.e., linear acceleration and angular velocity) inevitably include drift error from accumulation of sensor noise and bias. The IMU-based wearable system estimates the rotation of each joint firstly, then reconstructs the position of each joint by foward-kinematics. Thus, the erroneous link length of all segments also bring about accumulative error especially for the distal joints (i.e., the positions of finger-tips).

For the soft wearable hand tracking systems, the reconstruction of joint angle is normally estimated from the linear regression with the calibrated coefficients (e.g., $\theta = k_1\alpha + k_2$ where α and θ is raw soft sensor data and estimated joint angle respectively), thus the miscalibrated parameters lead to large tracking error. Moreover, since soft sensor is susceptible to unmodeled contact causing the deformation of the sensor, the aforementioned issues of the vision-based systems in manipulating other objects (e.g., grasping a bottle or equipping wearable devices) still exists.

Despite of these inevitable error of all tracking systems, fortunately, we human cannot precisely perceive the tracking error. Particularly for the VR applications with HMD (head-mounted display), where the users can only see the virtual world with the real world visual information completely blocked by the HMD, human only can perceive the true positions/configurations of their hand by proprioception. Since human proprioceptive perception is not perfect (Van Beers *et al.* 1998), this implies that a finger-tracking system would still be adequate if we can make its tracking error less than a certain detection threshold of visual-proprioceptive conflict of the users (Welch & Foxlin 2002).

Moreover, the effect of haptic feedback for finger-based interaction is also studied in this chapter. Since the improvements of not only task-performance, but also user experience by modifying the user perception are verified in many researches. We study particularly study the effect of cutaneous haptic feedback (i.e., utilizes motor actuation to provide contact feedback onto human finger-tip.) for perception of finger-based tracking error, which, we think, is most adequate form of haptic feedback for finger-based interaction. There have been researches (Kuchenbecker *et al.* 2008; Solazzi *et al.* 2010; Prattichizzo *et al.* 2010*b*) developed various designs of cutaneous haptic device to offer more realistic cutaneous sensation. This cutaneous haptic system is more promising for the consumer market as compared to, e.g., finger-based exo-skeleton haptic systems (e.g., (Fontana *et al.* 2013; Frisoli *et al.* 2005)), which are in general technically difficult to construct, and, therefore, usually too expensive to be commercially-viable. It has also been reported that, even with the absence of kinesthetic feedback, the cutaneous haptic feedback alone can often provide adequate haptic sensation for virtual manipulation, particularly when the magnitude of the required force feedback is not so large (Prattichizzo *et al.* 2012*b*; Jang & Lee 2014). Since human proprioceptive perception is not perfect (Van Beers *et al.* 1998), this implies that a finger-tracking system would still be adequate if we can make its tracking error less than a certain detection threshold of visual-proprioceptive conflict of the users (Welch & Foxlin 2002).

The goal of the study in this chapter is to answer the following two questions related to this visual-proprioceptive conflict for the wearable finger-based interaction (and with cutaneous haptics) via suitably-designed human subject psychophysics study: 1) what is the detection (absolute) threshold (Colman 2015) of the visual-proprioceptive conflict when performing finger-based operation in VR (or tolerable error between the real position of the finger (as proprioceptively perceived by the users) and its visual presence in the VR scene); and 2) if this detection threshold can be enlarged when cutaneous haptic feedback is used (i.e., cutaneous haptic feedback makes the visual presence of the finger more convincing, thereby, allowing for larger finger tracking error). Several results have been proposed on the visual-proprioceptive conflict, e.g.: static orientation error (Madsen & Stenholt 2014), latency and noise effect (Livingston & Ai 2008), and arm drive angle in virtual or mixed reality (Burns *et al.* 2005). It was also shown in (Holmes & Spence 2005) that the position perception of the human is determined by a weighted sum of visual, proprioceptive and other senses, suggesting that the haptic feedback can indeed affect the perception of finger-tracking error as also to be established in the current thesis. Similarly, the interplay between the haptic feedback and other sensory stimuli was also studied, e.g.: the role of haptic and vision in curvature perception (Drewing & Ernst 2006), pseudo-haptics by modifying visual cues (Jang & Lee 2014) and the effect of matching visual cues with haptic cues on position perception (Folegatti *et al.* 2009). Yet, to our knowledge, quantitative analysis of the visual-proprioceptive conflict for complex spatial motion and, further, quantitative analysis of the effect of haptic feedback on the visual-proprioceptive conflict has not been explored. In summary, we firstly identify the quantitative detection threshold and the role of haptic feedback on this conflict by means of human subject study in this chapter.

2.2. Experimental Environment

To perform human subject study of the visual-proprioceptive conflict and the effect of cutaneous haptic feedback, we utilize the experimental setup as shown in Fig. 2.1, which consists of motion capture system (MOCAP), HMD, and wearable cutaneous haptic feedback device on the

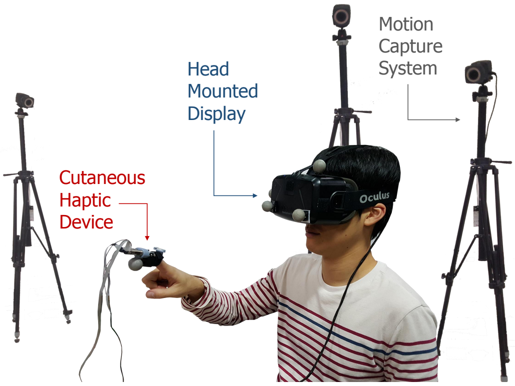


Figure 2.1. Experimental setup to study visual-proprioceptive conflict with cutaneous haptic feedback : human subjects wear cutaneous haptic device on their index finger and HMD, with their motions measured by motion capture system.

subject's index finger, each of them now detailed below.

2.2.1. *Hardware Setup*

In order to measure the motions of the human head and finger, we use VICON® MOCAP system, which provides the position and orientation of a set of reflective markers by using multiple IR cameras with 200Hz sampling rate and sub-millimeter spatial resolution. We attach a set of markers on the HMD to measure its pose (i.e., position and orientation). We also attach one marker as close as possible to the subject's finger-tip on the cutaneous haptic device to measure the finger-tip location.

All the experiments in this thesis are executed in the VR setting, i.e., all the visual information of the real world is completely blocked from the subjects. This is particularly crucial, because real visible hand will directly inform the subjects where their finger-tip is. To show only the virtual environment while also providing 3D immersive virtual visual information

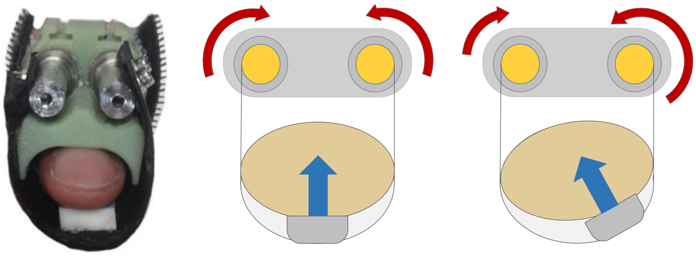


FIGURE 2.2. Cutaneous haptic feedback device (Jang & Lee 2014): normal force produced by rotating the two motors in opposite directions, while the shear force in the same direction.

generated for our experimental purpose, we adopt Oculus Rift® HMD, which provides 3-D vision with 1200 × 800 (640 × 800per eye) resolution and 90 degree field of view, which is one the widest among commercial HMDs.

For generating haptic feedback on the user's finger-tip, we utilize a cutaneous haptic feedback device as shown in Fig. 2.2, which was first proposed in (Minamizawa *et al.* 2007) and later adopted in many researches (e.g., (Prattichizzo *et al.* 2010*b*; Jang & Lee 2014)). Our cutaneous haptic device has two motors (Maxon DCX motor, $\phi = 10$mm, 3W, 16:1 gear ratio) with encoder providing the control axis with the resolution of 1024 count/rev. The motors are connected to a desktop PC via US Digital® USB4 DAQ board and Arduino® board, with the sampling rate of about 1kHz. The rubber block attached between the finger-tip and the band is manipulated by the motors. We can then transmit normal or shear force by controlling the two motors to their respective designated angles - see Fig. 2.2.

2.2.2. *Virtual Environment Rendering*

The virtual environment for our experiments is constructed using the above equipments and OpenGL®. In the virtual environment, the virtual sphere, which represents the position of the index finger-tip is shown to the subjects by measuring the relative position of the HMD and the finger-tip by using the MOCAP system. The subjects are noticed that the marker on the cutaneous haptic device will represent their finger-tip, thus, they should consider this marker as their finger-tip position throughout all the experiments. The radius of this finger-tip sphere is set to 1cm which is the same as the radius of the marker. During the experiments, the human subjects can freely move their hands and head.

To generate the contact force with this finger-tip sphere, we create a virtual half cylinder fixed in the virtual environment as the contact target for task of following experiments as used in (Drewing & Ernst 2006). The reason why we choose this cylindrical shape is that the contact force with this shape can be fully implemented by combination of shear and normal force which can be generated by our 2-DOF cutaneous haptic device. The radius and height of the cylinder is set to be 25cm and 50cm respectively. This size of the cylinder is chosen to accommodate the human motion during the human subject study experiments in following sections. The details of generating contact force will be explained in Sec. 2.4. See Fig. 2.3 for how the subjects see the virtual environment, the finger sphere and the half cylinder in the HMD.

2.2.3. *HMD Calibration*

If we directly measure the pose of the real finger-tip w.r.t. the pose of the real HMD by using the MOCAP system, and render it to the user via

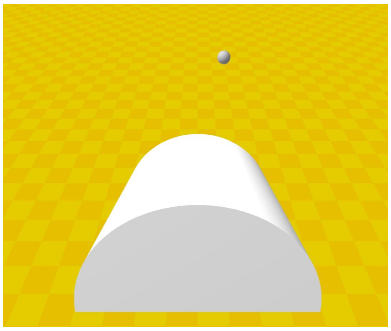


FIGURE 2.3. Virtual environment as seen from the HMD with the sphere of user index finger-tip position and the half-cylinder for performing contact task.

the HMD in the HMD frame $\{\mathcal{H}\}$ as measured by using MOCAP system, we find the rendered image appears often too close to the human eyes than how it should appear in the real world. This we think is because the graphics of the HMD, as perceived by the user, is not rendered w.r.t. the physical HMD frame $\{\mathcal{H}\}$, but rather w.r.t. another abstract frame, which we denote by $\{\mathcal{G}\}$ and call it *graphics rendering frame*. See Fig. 2.4. In order to reduce this pose-difference between the HMD frame $\{\mathcal{H}\}$ and the graphics rendering frame $\{\mathcal{G}\}$, we perform the HMD calibration as follows.

First, we assume that the main factors dictating the distortion of the graphics rendering in the HMD is related to the translation and orientation of $\{\mathcal{G}\}$ w.r.t. $\{\mathcal{H}\}$ because of the kinematic discrepancy between the HMD frame (i.e., randomly attached MOCAP markers) with the exact graphic rendering frame. We can then relate the two frames $\{\mathcal{H}\}$ and $\{\mathcal{G}\}$ by a rigid body transformation with the relative translation offset $p_{\mathcal{G}}^{\mathcal{H}} \in \Re^3$ and the relative rotation offset $R_{\mathcal{G}}^{\mathcal{H}} \in \mathrm{SO}(3)$ from $\mathcal{G}$ to $\mathcal{H}$. Further, if we denote by $p_f^{\mathrm{cmd}} \in \Re^3$ the graphics rendering command of a point p_f in the HMD, what we would have is:

$$R^{\mathcal{H}}_{\mathcal{G}} p^{\mathrm{cmd}}_f + p^{\mathcal{H}}_{\mathcal{G}} = p^{\mathcal{H}}_f \tag{2.1}$$

where note that p^{cmd}_f is applied to the spot where $p^{\mathcal{G}}_f$ should be, as, here, again, we assume the graphics rendering reference frame is located at $\{\mathcal{G}\}$ not $\{\mathcal{H}\}$ based on our observation.

Now, in order to estimate $p^{\mathcal{H}}_{\mathcal{G}}$ and $R^{\mathcal{H}}_{\mathcal{G}}$, we perform the following *reaching-without-seeing task*, which utilizes human proprioception as a sensing mechanism. More precisely, we randomly generate a virtual sphere at $p^{\mathrm{cmd}}_f \in \Re^3$ and ask the subjects to move their finger-tip to this rendered sphere as close as possible. During this task, only the virtual sphere is rendered as a gray sphere, while the user finger-tip is not rendered (i.e., they move their finger-tip only by relying on their proprioceptive perception). This subject's finger-tip position can then be measured by the MOCAP system w.r.t. the physical HMD frame $\{\mathcal{H}\}$, which can be written by $p^{\mathcal{H}}_f \in \Re^3$.

We can then utilize (2.1) to identify $p^{\mathcal{H}}_{\mathcal{G}}$ and $R^{\mathcal{H}}_{\mathcal{G}}$, since p^{cmd}_f is the known command and $p^{\mathcal{H}}_f$ is measured by the MOCAP system. In fact, (2.1) reduces to a linear equation of $p^{\mathcal{H}}_{\mathcal{G}}$ and $R^{\mathcal{H}}_{\mathcal{G}}$ in this case. Thus, by repeating this reaching-without-seeing task many times, we can produce a data set of $p^{\mathrm{cmd}}_f, p^{\mathcal{H}}_f$, and, using this data set and (2.1), we can estimate $R^{\mathcal{H}}_{\mathcal{G}}, p^{\mathcal{H}}_{\mathcal{G}}$, thereby, completing the HMD calibration.

For this calibration and also throughout the following experiments, we utilize the spherical coordinate system (r, θ, ϕ) as shown in Fig. 2.4. The radius r, the azimuth θ and the elevation ϕ of the virtual sphere are randomly chosen from the following ranges: $\theta, \phi \in [-30^\circ, 30^\circ]$ and $r \in [30, 50]$cm. These ranges are chosen, as the majority of our experiments

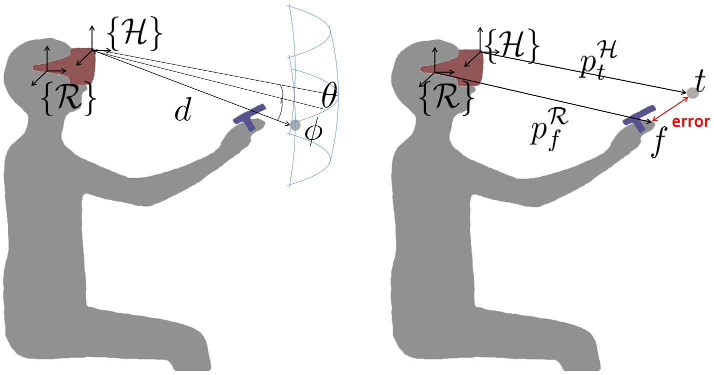


FIGURE 2.4. HMD frame $\{\mathcal{H}\}$ and graphics rendering frame $\{\mathcal{G}\}$: if graphics is rendered w.r.t. $\{\mathcal{H}\}$ as defined by HMD manufacturer, in reality, it is rendered w.r.t. abstract frame $\{\mathcal{G}\}$. We also use spherical coordinate (r, θ, ϕ) attached at $\{\mathcal{H}\}$ throughout this thesis.

(and other typical tasks in the VR with the HMD) is performed around their corresponding regions. Five subjects participate in the calibration task and take 100 trials each, making total 500 data points. We then use the results in (Besl & McKay 1992) for obtaining the least square solution of the rigid-body transformation.

2.3. Identifying the Detection Threshold of Tracking Error

The Experiment #1 is "Identifying the Detection Threshold of Tracking Error", the purpose which is to measure the detection threshold of the tracking error (i.e., visual-proprioceptive conflict) human can perceive in the virtual environment while moving their finger without haptic feedback. We measure this threshold by asking subjects to differentiate the measured position of their finger-tip by using the MOCAP system (i.e., true finger sphere) from the one intentionally perturbed with some position error from the measured finger position (i.e., false finger sphere).

2.3.1. *Experimental Setup*

Eight human subjects participate in the Experiment #1. They are all male, from the age of 22 to 32, right-handed with no known perception disorder and use their index finger of dominant hand for this experiment.

The virtual environment, displayed to participants through the HMD, primarily consists of the the true and false finger spheres and the half cylinder as mentioned in Sec. 2.2.2. The half cylinder is also rendered horizontally with its center locating at $(r, \theta, \phi) = (45cm, 0°, -60°)$ and extending toward far from the subject. During the Experiment #1, the cutaneous haptic device is turned off to exclude the effect of haptic feedback. In contrast, Experiment #2, detailed in Sec. 2.4, is performed with this haptic feedback turned on, and, to minimize any bias stemming from the order of these two experiments, we randomly alter between them for each subject after the initial familiarization phase as explained below, resulting in the 4 subjects starting with the Experiment #1 and the other 4 subjects with the Experiment #2.

2.3.2. *Procedure*

Before performing Experiment #1, the subjects are given enough time to be familiar with the virtual environment to minimize learning effect during the experiment (i.e., initial familiarization phase). They are able to move their finger-tip freely in empty space or touch the cylinder with no haptic feedback. They then proceed to Experiment #1, which consists of the following three phases as illustrated in Fig. 2.5: **adaptation phase**, **blind phase** and **answer phase**.

In the **adaptation phase**, the subjects are asked to swipe the surface of the cylinder from side to side. This adaptation phase lasts for 5 seconds

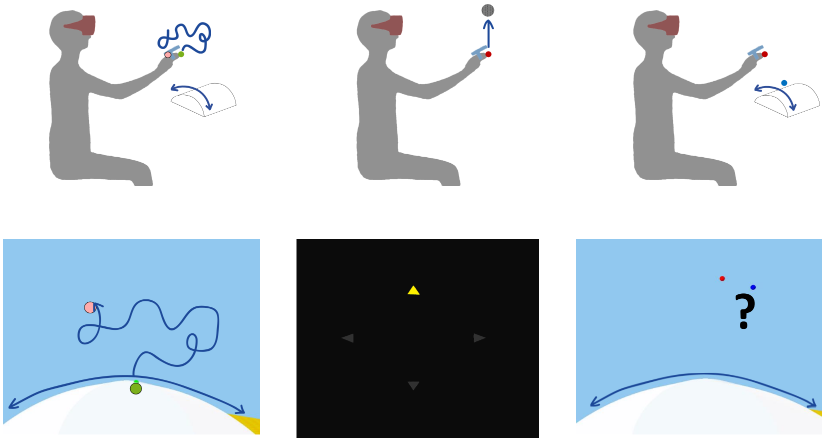


FIGURE 2.5. Experiment #1: (1) **adaptation phase** (left): subjects freely swipe the surface of cylinder; (2) **blind phase** (middle): subjects move their finger following yellow arrow with the virtual environment blacked out; (3) **answer phase** (right): subjects answer which sphere corresponds to their real finger-tip after the cylinder swiping task with true and false finger spheres.

to reaffirm the mapping between the visual cue in the virtual environment and the proprioception in the real world of the subject, which is acquired during the pre-experiment familiarization phase as stated above. During this phase, the false finger sphere is not rendered.

In the **blind phase**, which lasts for 3 seconds, all the virtual environment and the true finger sphere are blacked out. Instead, in the black screen of the HMD, an arrow is rendered in a pre-defined position to inform the subject of one of the four directions - up, down, left and right. See Fig. 2.4. This direction of the arrow is updated every 1 second based on the location of the subject's real finger-tip (not shown to them) in such a way to minimize the distance from the center line of the HMD screen (i.e., $\theta = \phi = 0$), but not strictly enforcing it, so that the subject can start the next answer phase with their finger suitably positioned. The main role of this blind phase is to separate the adaptation phase and the answer phase by nullifying the subject's perception of the true finger sphere during the adaptation phase.

The virtual environment and the true finger sphere appear again in the **answer phase** similar to the case of adaptation phase. However, in the answer phase, the false finger sphere is also rendered, whose position is defined by perturbing that of the true finger sphere with randomly generated error. More precisely, we define a sphere, whose center is at that of the true finger sphere and whose radius is randomly chosen from the range from 1.5cm to 7.5cm with the interval of 1.5cm. This range is chosen based on a pilot experiment, in which we found the detection threshold would exist within this range. The center of the false finger sphere is then (continuously) randomly generated on the surface of the sphere with this randomly-chosen radius and centered at that of the true finger sphere - see Fig. 2.5. The color of the true and false fingers is also chosen randomly between blue and red for each trial of the experiment. Once this relative vector of the false finger sphere is set from the true finger sphere, the two spheres are rendered to move together under the subject motion command, except when one (or both) of them makes contact with the cylinder, during which the penetration of each of the spheres into the cylinder is graphically removed (see also Fig. 2.7).

The human subjects are then asked to carry out the classical Two Alternative Forced Choices (2AFC) with those two spheres, while moving the spheres and making contact on the surface of the cylinder with them. For each trial, the subjects are instructed to consider one of the two spheres as if it is their real finger-tip and perform the swiping task. They

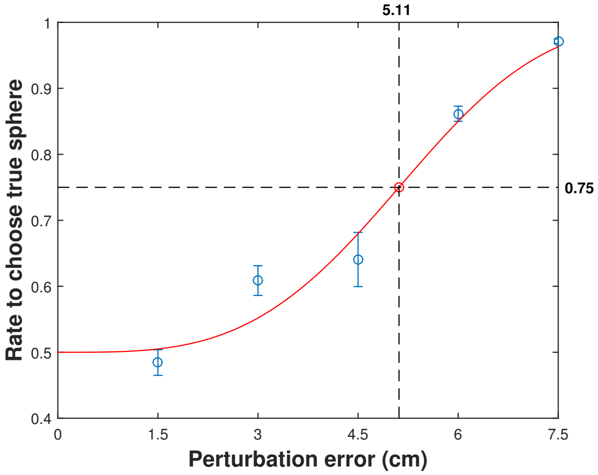


FIGURE 2.6. Result of Experiment #1: mean and variance of the average rate of each subject choosing the true finger sphere over the false finger sphere (fitted by the psychometric curve).

are then asked to repeat this task while considering the other sphere (with different color) as their real finger-tip position. After each trial of this task, the subjects are asked to choose which finger sphere (i.e., blue or red) represents the true position of their finger-tip. The subjects are told to make their decision as soon as possible, but also told that making a correct decision is more important. If they say they need more trials to answer the question, we allow them to do so. For each subject, 8 trials (or answers if the subject requested more trials) of the above 2AFC test are performed with each of the 5 error steps between the true and false finger spheres, resulting in the total 40 answers of 2AFC of each subject for the Experiment #1.

2.3.3. *Experimental Result*

The mean and variance of the average correct answer rate of each subject choosing the true finger sphere over the false finger sphere are plotted by the blue markers and lines in Fig. 2.6. As expected, the rate of correct answer of choosing the true finger sphere increases as the perturbation error becomes large. The detection threshold of this visual-proprioceptive conflict is then defined to be the value where the correct choice occurs with the rate of 75% (Allin *et al.* 2002). To estimate the tracking error corresponding to this detection threshold, we fit the obtained data to the Weibull function (Wichmann & Hill 2001). We then obtain the detection threshold value to be 5.11cm, when the operation takes place about 30cm away from the subject's eyes (i.e., w.r.t. $\{\mathcal{G}\}$ in Fig. 2.4).

2.4. Enlarging the Detection Threshold of Tracking Error by Haptic Feedback

The Experiment #2 is "Enlarging the Detection Threshold of Tracking Error by Haptic Feedback", the main objective of which is to quantitatively analyze the effect of the cutaneous haptic feedback on the detection threshold of the visual-proprioceptive conflict as obtained in Experiment #1 of Sec. 2.3.

2.4.1. *Experimental Setup*

The set of participants is the same between Experiment #1 and Experiment #2. The order of Experiment #1 and #2 is counterbalanced.

The setting of the experiment is also the same as that of Experiment #1 in Sec. 2.3, except that the cutaneous haptic feedback is activated. To mitigate bias associated with the order of the trials with or without

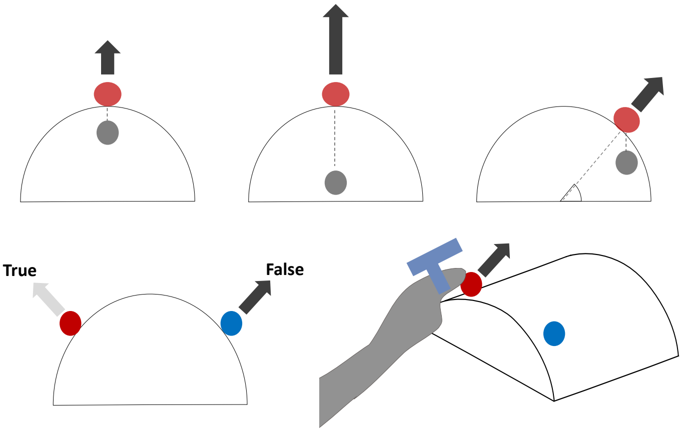


Figure 2.7. Haptic feedback generation for Experiment #2: (top) when the sphere penetrates the cylinder (gray sphere), haptic feedback is generated using the vertical distance from surface, whereas penetration removed from graphics shown to human subjects (red sphere); (bottom) during the answer phase, subject perceives haptic feedback of the randomly-chosen sphere, not necessarily same as true finger sphere as illustrated here.

the haptic feedback, we randomly select Experiment #1 and Experiment #2 for each subject after the initial familiarization stage as explained in Sec. 2.3.

2.4.2. *Procedure*

The same procedure as in Experiment #1 is used for Experiment #2, except that the cutaneous haptic feedback is turned on during the adaptation phase (for the true finger sphere) and during the answer phase (for the randomly-chosen true or false finger sphere). When the sphere makes contact with the cylinder surface, the corresponding contact force is generated and conveyed to the subject as a combination of the normal and shear forces as shown in Fig. 2.7. For this, we use the initial magnitude

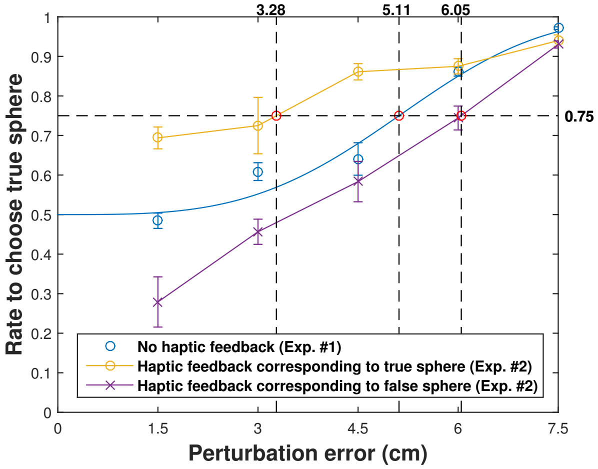


FIGURE 2.8. Result of Experiment #2: mean and variance of average rate of each subject correctly identifying the true finger sphere with haptic feedback either on true or false sphere.

force , when contact begins, as 0.45N and the stiffness $K = 0.042\text{N/cm}$ for the contact force generation, to simulate within the range of our haptic device's force generation capability. Since the human subjects do not know where the haptic feedback is generated from, either from the true or false finger sphere, we can investigate how the correct (i.e., randomly-chosen sphere = true finger sphere) or false (i.e., randomly-chosen sphere = false finger sphere) cutaneous haptic feedback affects the human detection threshold of the visual-proprioception conflict. The same number of 2AFC as for the Experiment #1 for each subject is also performed in this Experiment #2. See also Fig. 2.7 for depiction of some haptic feedback generation procedures for Experiment #2.

2.4.3. *Experimental Result*

The results of Experiment #2 are shown in Fig. 2.8 with the curve in Fig. 2.6 from Experiment #1 also shown there, where the x-axis is the magnitude of perturbation error between the true and false finger spheres. Then, we can see from Fig. 2.8 that, with the haptic feedback on the true finger sphere, the rate of correctly identifying the true finger sphere increases as compared to the case of no haptic feedback (i.e., middle line in Fig. 2.8). On the other hand, with the haptic feedback on the false sphere, the rate of correctly identifying the true sphere decreases as compared to the case of no haptic feedback. We can further compute the detection threshold for the cases of correct and false haptic feedback as done for the Experiment #1 by fitting their data as shown by the top and bottom curves in Fig. 2.8. For this, we use linear interpolation instead of using Weibull function as done for the Experiment #1, since the two curves necessitate the violation of typical form of the Weibull function with the rate required to converge to 50% with no stimuli. Based on the linear interpolation, we can then obtain the detection threshold of visual-proprioceptive conflict to be 3.28cm for the case of correct haptic feedback (i.e., haptic feedback on true finger sphere) and 6.05cm for the case of false haptic feedback (i.e., haptic feedback on false finger sphere).

2.5. Discussion

From Fig. 2.8, we can see that: 1) with the haptic feedback on the true finger sphere, the rate of correctly identifying the true finger sphere increases, which we believe is because the correct haptic feedback would re-enforce the human's correct perception of the true finger of the case with no haptic feedback (i.e., middle curve of Fig. 2.8); and 2) with the haptic feedback on the false sphere, the rate of correctly identifying the true sphere decreases, suggesting that the false haptic feedback can confuse human subjects, thereby, enlarging their detection threshold (i.e., becoming less sensitive) of the visual-proprioceptive conflict.

This is quantitatively evidenced by their respective detection thresholds, that is, as compared to the detection threshold 5.11cm for the case of no haptic feedback (i.e., middle curve in Fig. 2.8), that for the case with the correct haptic feedback decreases to 3.28cm (i.e., top curve in Fig. 2.8), whereas that with the false haptic feedback increases to 6.05cm. This then means that, for VR applications with no haptic feedback, human users would likely perceive the rendered finger-tip (i.e., false finger sphere) as their real finger-tip (i.e., true finger sphere) not any more, if the finger tracking error becomes larger than 5.11cm, whereas this tolerable tracking error will be increasing by about 1cm if the VR applications involve cutaneous haptic feedback. On the other hand, with the cutaneous haptic feedback on the real finger-tip (i.e., true finger sphere), it would be much more difficult to "fool" the human users to believe the rendered sphere (i.e., false finger sphere) as their real finger-tip, as evidenced by the sharp decrease (i.e., about 2cm) of the detection threshold from the middle to the top curves in Fig. 2.8.

More precisely, consider the bottom curve of Fig. 2.9, which is more useful for our purpose here as it specifies how probable human users would perceive the rendered avatar (i.e., false finger sphere) when haptic feedback is synchronized with its motion. The indicator, which may be most useful to design the tracking performance specification, would then be the error with 50% rate of subjects choosing the false sphere as their real finger-tip (i.e., 3.50cm in Fig. 2.9). This is because, if the finger track-

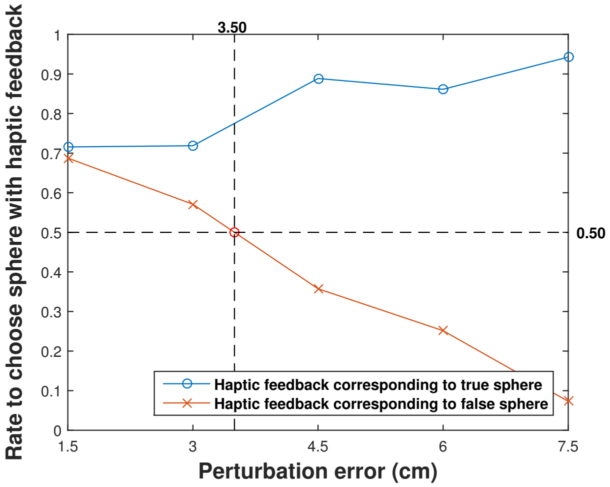


FIGURE 2.9. Rate of subjects choosing the (true or false) sphere with the haptic feedback as their real finger-tip position.

ing performance error is less than this value, the human users would not be able to distinguish the rendered finger sphere from their real finger sphere, defining minimum performance for the tracking system. Due to this reason, we call this 50% rate point of the curve in Fig. 2.9 *maximum allowable tracking error* for the finger-based haptically-enabled VR applications. Note from Fig. 2.9 that, to be useful for our finger-based haptic system setting, the tracking system should at least perform with its tracking error less than 3.50cm.

To closely investigate the effect of cutaneous haptic feedback, we obtain the graph in Fig. 2.9 from Fig. 2.8, which shows the rate of subjects choosing the sphere with the haptic feedback as their real finger-tip, although the haptic feedback may be exerted on the false (or true) finger sphere. From Fig. 2.9, we can then see that, the smaller the perturbation error is, the more dominant the role of the haptic feedback is on the human's perception on their finger-tip location, as evidenced by that about

70% of the subjects choose the (true or false) sphere as their real finger-tip as long as it has the haptic feedback (i.e., relying more on the haptic feedback than proprioceptive perception). This is in fact true both for the cases of the correct and false haptic feedback, breaking the 50% rate of choosing either of the sphere for the case of no haptic feedback (see Fig. 2.6). As the visual-proprioceptive error becomes larger (i.e., larger perturbation error), the dominance of the haptic feedback seems to cede to the proprioception, as can be seen from the rapid decrease of choosing the false finger sphere as the real finger-tip position, even if the haptic feedback is still on the false sphere. This changing role of the cutaneous haptic feedback depending on the perturbation error is also reaffirmed by one way ANOVA test, for which we choose the type of haptic feedback (i.e., no, correct and false haptic feedback) as independent factor and the answer rate of subjects correctly identifying the true sphere as the dependent variable. Then, in the case of perturbation error 7.5cm, no significant difference is found ($F_{2,15} = 1.6, p > 0.05$), while significant difference is found for the perturbation error 1.5cm ($F_{2,15} = 6.9, p < 0.0001$).

Chapter 3

Wearable Finger Tracking Module for Haptic Interaction

3.1. Motivation

The efficacy of haptic feedback for finger-based interaction has been studied and verified in recent years (Meli *et al.* 2018; Bimbo *et al.* 2017; Baldi *et al.* 2015; Gleeson *et al.* 2010; Scheggi *et al.* 2015). The result of the previous chapter (i.e., enlarging the detection threshold by haptic feedback) also would be one of great examples, which alleviate the error perception of finger-tracking system via haptic feedback. In this context, wearable haptics for VR has received great attention and been under active investigation by many research groups and companies around the globe. Among many forms of wearable haptics, particularly promising is the multi-finger-based wearable haptics, since it allows for the VR realization of many real-life scenarios and interactions, which typically involve heavy usage of fingers and hands. In fact, using the fingers and hands is argued as one of the key characteristics of our human being itself, thus, we believe that finger-based wearable haptics is crucial to attain truly

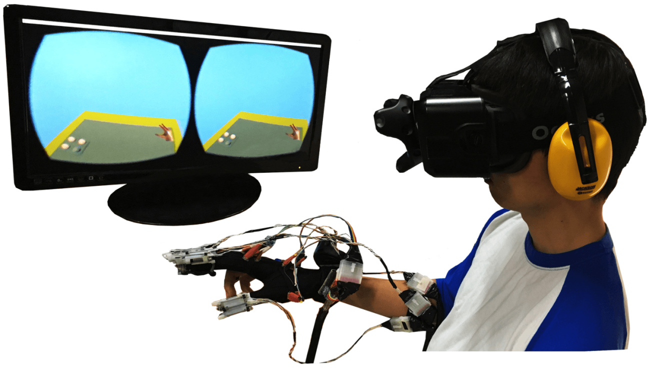


Figure 3.1. Multi-fingered virtual manipulation with the proposed finger tracking module (FTM) integrated with cutaneous haptic device (CHD): a peg insertion task into a horizontally-placed hole.

immersive, multifarious and real-life like VR experiences. Especially the cutaneous haptic devices (CHDs) are widely adopted and researched as its smaller form-factor and structure suitable for finger-based interaction than kinesthetic haptic devices (Pacchierotti *et al.* 2012; Jang & Lee 2014).

For this finger-based wearable haptics for VR, the primary requirement is reliable tracking of multiple fingers and hands in various motions and postures as mentioned in the previous chapter. Many methodologies have been proposed for this multi-finger tracking. Vision-based technology (e.g.,(Oikonomidis *et al.* 2011; Kim & Park 2015; Meli *et al.* 2014; Maisto *et al.* 2017)) has attracted many researchers for tracking the hand with haptic devices. However, due to the visual distortion from additional haptic devices equipped on the user's hand (e.g., kinesthetic haptic devices (Wang *et al.* 2018; Prachyabrued & Borst 2015), cutaneous haptic devices (Pacchierotti *et al.* 2012; Chinello *et al.* 2017; Perez *et al.* 2016), vibrotactile haptic devices (Maereg *et al.* 2017)), vision-based tracking often fails to track the fingers when some extra devices are attached on them. This is because, the vision-based method, which dominantly are based on machine-learning-based methods, would not properly work under conditions (i.e., generalization issues). This issue is also well addressed in the recent hand tracking challenge (Armagan *et al.* 2020), which give a few tasks to verify how well generalized a vision tracking algorithm is in terms of four generalization factors; individually different hand shape (size, color, texture, etc.), articulation (hand gesture), camera viewpoint, and object. A fair number of state-of-the-art vision-based methods show miserable performance and even winner of each task (Zhang *et al.* 2020; Iqbal *et al.* 2018) shows fairly dropped accuracy (about 2-4 times scaled error) for untrained conditions, which can clearly imply the fact that the current-available vision-based method based on bare hand cannot be compatible with most of haptic devices (Lee *et al.* 2019; Baldi *et al.* 2017), even, with any extra devices (e.g., wearing glove for the disabled (In *et al.* 2015; Kim *et al.* 2019), sports, or safety). In this context, most of current vision-based tracking systems are trained from a vast dataset of bare human hand (Mueller *et al.* 2018; Zimmermann & Brox 2017), thus, the visual distortion of human hand induced by attaching extra wearable haptic devices would largely degrade the performance of hand tracking.

IMU-based system (with compass) would be alternative way to be compatible with haptic devices partially. However, this type of system also has issue of magnetic-interference (i.e., instability near the magnetic objects) for compass as mentioned in the introduction. The problem when integrating the CHDs with IMU-based tracking system is the fact that the cutaneous haptic devices normally has dc-motors or servo-motors for generating the force feedback or the finger-tip. The embedded motors or their operating current surely disturb the magnetism reading of the compass, which is the reason why IMU-based tracking system has difficulty to integrate with CHDs purely.

Along this reasoning, in this chapter, we propose a novel glove-type **finger tracking module (FTM)**, which opportunistically utilizes IMU sensors and soft sensors to estimate multi-DOFs finger/hand motion while being free from the electromagnetic interference issue of the IMUs and the complex sensor wrapping/arrangement issue of the soft sensors. The proposed FTM is designed in the simple glove-form for easy integrated (or separate) usage with haptic devices and straightforward implementation. We determine the attaching locations of each sensor to minimize the system complexity while also carefully observing the finger/hand anatomy different sensing capabilities and characteristics of each sensor. Our proposed FTM also only requires a simple three-step known-pose calibration.

We combine this FTM with the CHDs improved from our previous work (Jang & Lee 2014), which can generate high-performance three-DOF finger-tip haptic feedback by utilizing miniatured DC motors with accurate feedback control with FSR and soft sensors. By combination of these FTM and CHD, we construct wearable cutaneous haptic interface (WCHI) as shown in See Fig. 3.1, which can facilitate dexterous/immersive multi-finger haptic interaction. The versatility of our proposed FTM for haptic interaction is verified in this chapter. A few researches study about the way to finger tracking with extra haptic devices (i.e., accurate finger tracking while delivering haptic feedback), however, 1) the device of (Weber *et al.* 2016) utilizes only a single IMU on the dorsum of the hand and a single soft sensor for each finger, thus, not able to fully track large-DOF complex finger motions; 2) Commercial products (e.g., MANUS VR ®) glove can fully track the thumb motion, yet, not the adduction-abduction (AA) motion of index/middle fingers, which turns out to substantially affect VR experience, particularly it involves complex/dexterous finger motion (see Sec. 3.3.3); and 3) both of these devices provide only single-DOF vibro-tactile feedback on the finger-tip, too simple to capture most of real-life finger/hand interactions, and which are free from catastrophic magnetic-interference problem when adopting more dexterous/accurate cutaneous haptic modules. In contrast to this, our proposed WCHI can fully track complex/dexterous large-DOF finger/hand motions including the finger AA motion, while also providing three-DOF cutaneous finger-tip haptic feedback. The FTM is also designed in such a way that they can easily integrated with the CHD into the WCHI without mechanical and functional interferences or used separately.

3.2. Development of Finger Tracking Module

3.2.1. *Hardware Setup*

To design our finger-tracking module compatible with the haptic device, the key issues of vision-based hand tracking are summarized again that: 1) the large-DOF finger motion within a small region, thus, such commercial systems as HTC VIVE, Kinect or VICON, which can fairly well track "larger" arm or wrist motions, cannot be directly used; and 2) the dexterous finger motion, combined with omni-directional wrist motion, frequently induces the issue of occlusion, which is fundamental for any vision-based systems (e.g., LeapMotion) and has not yet been overcome. Due to these reasons, in this thesis, we aim to develop FTM (finger tracking module) with IMUs and soft sensors, all attached in the glove

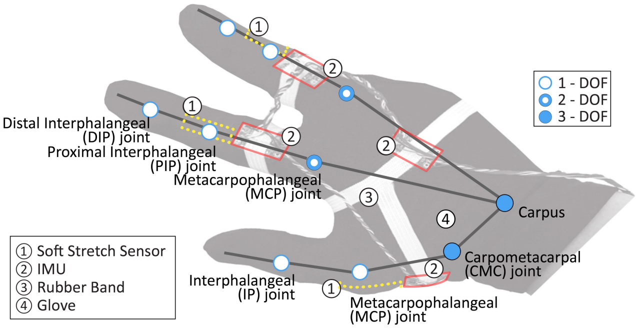

①
②
①
②
②
○ 1 - DOF
◉ 2 - DOF
● 3 - DOF
Distal Interphalangeal (DIP) joint
Proximal Interphalangeal (PIP) joint
Metacarpophalangeal (MCP) joint
③
④
Carpus
Carpometacarpal (CMC) joint
① Soft Stretch Sensor
② IMU
③ Rubber Band
④ Glove
Interphalangeal (IP) joint
①
②
Metacarpophalangeal (MCP) joint

FIGURE 3.2. Finger tracking module (FTM) consists of IMUs and soft sensors embedded in the form of glove (hand model with links and joints also illustrated).

form, so that large-DOF/small-size finger motion can be tracked while avoiding the issue of occlusion. See Fig. 3.2, where we assume the wrist position information is provided by a commercial external vision sensor (e.g., HTC VIVE) and also anatomical constants (i.e., link length, joint position, etc.) given from off-line identification (Chang & Pollard 2008). The FTM fully tracks each segment of the thumb, index and middle fingers. To determine which sensor is attached to which segment, we carefully consider the DOF of each joint, which varies from one to three.

More precisely, we decide the joints, the finger segments and the sensor arrangement for the FTM as shown in Fig. 3.2, where: 1) hand dorsum/carpus is with three-DOF rotation (e.g., wrist rotation); 2) three-DOF carpometacarpal (CMC) joint with FE (flexion-extension), PS (pronation-supination) and AA (abduction-adduction) motions; and 3) two-DOF metacarpophalangeal (MCP) joint of the index/middle fingers with FE and AA motions. To estimate the (relative) orientations of these joints with a single sensor attachment, we attach four MEMS IMUs (InvenSesne® MPU9250) on the dorsum of the hand, on the first metacarpal of the thumb, and the proximal phalanges of the index and middle fingers, re-

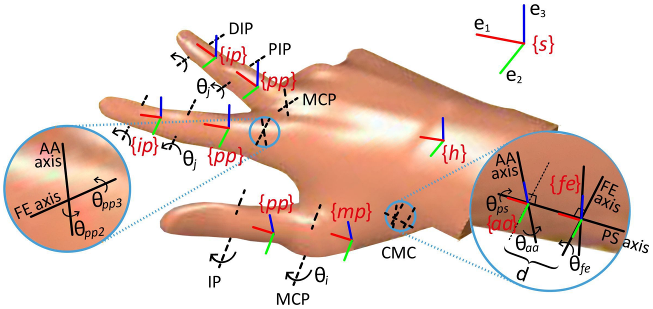


FIGURE 3.3. Hand model with the joints, links, and coordinate frames: thumb possesses three-DOF CMC joint, single-DOF MCP and IP joints; whereas index and middle fingers each possesses two-DOF MCP joint, single-DOF PIP and IP joints.

spectively. We choose these IMUs over soft sensors here, since: 1) they can provide three-DOF global rotation information at once; and 2) their attachment point is more flexible than soft sensors, that must be attached wrapping over the joint. These IMUs are fastened by rubber bands in the form of glove. We also attach the IMUs as far from the finger-tip CHD as possible to avoid electromagnetic interference between the motors of CHMs and the magnetometers of the IMUs (see Sec. 3.3 for interference test result). On the other hand, we utilize three capacitive type soft sensors (StretchSense™) and embed them inside the glove of each finger, to estimate the single-DOF bending angles of the thumb MCP joint and the proximal interphalangeal (PIP) joints of the index/middle fingers. Here, we consider the thumb MCP joint to be single-DOF with FE motion, since its AA motion caused by its inter-connection with the CMC joint is relatively small (Hollister *et al.* 1995; Kim *et al.* 2002) especially during the finger grasping and manipulation.

In addition, we utilize the musculoskeletal dependency (i.e., synergy) to estimate the motion of the thumb interphalangeal (IP) joint and distal interphalangeal (DIP) joints of the index/middle fingers from the thumb MCP joint and index/middle finger PIP joints. Note that we utilize the synergy not only for index/middle fingers but also for the thumb. This synergy was investigated in (Hrabia *et al.* 2013) and later employed in (Baldi *et al.* 2015). Hrabia et al. (Hrabia *et al.* 2013) showed that the synergy of the thumb MCP joint ($R^2 = 0.59$) is weaker than that of index finger's DIP joint ($R^2 = 0.77$), yet, still similar to that of little finger ($R^2 = 0.63$). In this chapter, we adopt this thumb synergy with its strength determined by trial-and-error, which turns out to be adequate for our purpose, that is, the FTM for VR applications, where believable graphics and haptic sensations are enough as evidenced/illustrated through our (rather extensive) experiments (see Sec. 3.3) in contrast to, e.g., medical applications, where accuracy is more weighted. Utilizing the soft sensors and the synergies for single-DOF joints and IMUs for multi-DOF joints, we can reduce the number of the sensor attachments, resulting in simpler estimation algorithm and lesser electromagnetic interference when integrated with the CHD. In addition, we employ five-DOF thumb model, which represent each thumb joint as single-DOF (IP, MCP) or three-DOF (CMC).

3.2.2. *Tracking algorithm*

In order to reconstruct the poses of three fingers and hand from FTM, the forward kinematics is applied to each joint of the fingers and the hand. Let $p^s_{s,h} \in \Re^3$ be the position vector from the origin of the inertial frame $\{s\}$ to that of the hand frame $\{h\}$ expressed in $\{s\}$-frame, where $\{h\}$-frame is attached to the hand dorsum as shown in Fig. 3.3. In this thesis, we use an external low-cost vision sensor (e.g., HTC VIVE tracker) to measure this $p^s_{s,h}$. Denote the pose of the $\{h\}$-frame relative to the $\{s\}$-frame by the homogeneous transformation $\bar{g}^s_{s,h}$ $\in$SE(3), i.e.,

$$\bar{g}^s_{s,h}(R^s_{s,h}, p^s_{s,h}) = \begin{bmatrix} R^s_{s,h} & p^s_{s,h} \\ 0 & 1 \end{bmatrix} \in \mathrm{SE}(3)$$

where the $R^s_{s,h} \in \mathrm{SO}(3)$ is the rotation of $\{h\}$ w.r.t. $\{s\}$, which is to be measured by the IMU/compass attached to the $\{h\}$-frame as shown in Fig. 3.3.

More specifically, we adopt a singular value decomposition method (Markley 1988) to provide prior estimation of the IMU/compass. The sensor data of accelerometer and compass in current body frame $\{h\}$ is compared with the reference sensor data, which represent the gravity and earth-magnetic field direction in inertial frame $\{s\}$. This rotational estimation, yet, only is valid for low frequency signals, since the high frequency erroneous factors such as IMU noise or linear acceleration are reflected in the estimation. Thereby, the gyroscope, which can provide accurate information of high frequency motions in the form of angular rate, should be integrated with the prior estimation, through the estimation algorithm such as complementary filter, extended Kalman filter. We also adopt a nonlinear complementary filtering algorithm (Mahony *et al.* 2008) to fuse the low frequency estimation (accelerometer/compass) and high frequency estimation (gyroscope) opportunistically.

From the above rotational estimation results, the methods to estimate motions of each finger by forward-kinematics are different from the case of thumb finger motion and the cases of index/middle finger motion. For the thumb motion, we attach the frames $\{fe\}$ and $\{aa\}$ to the CMC joint and $\{mp\}$ to the metacarpal bone between the MCP and CMC joints to respectively express the FE motion θ_{fe}, the PS motion θ_{ps}, and the AA motion θ_{aa} of the CMC joint with the offset among their axes also taken into account - see Fig. 3.3. We then have the following kinematics of the $\{mp\}$-frame expressed in the $\{s\}$-frame:

$$\bar{g}^s_{s,mp} = \bar{g}^s_{s,h} \cdot \bar{g}^h_{h,mp}(R^h_{h,mp}, p^h_{h,mp}) \in \mathrm{SE}(3) \tag{3.1}$$

with

$$\begin{aligned} R^h_{h,mp} &= R^h_{h,fe} R^{fe}_{fe,aa} R^{aa}_{aa,mp} = R^{s,T}_{s,h} R^s_{s,mp} \\ p^h_{h,mp} &= p^h_{h,fe} + R^h_{h,fe} p^{fe}_{fe,aa} + R^h_{h,aa} p^{aa}_{aa,mp} \end{aligned} \tag{3.2}$$

where $R^h_{h,fe} = \exp(\theta_{fe} e_2)$, $R^{fe}_{fe,aa} = \exp(\theta_{ps} e_1)$, and $R^{aa}_{aa,mp} = \exp(\theta_{aa} e_3)$ with the corresponding frames initially aligned with each other, $R^h_{h,aa} = R^h_{h,fe} R^{fe}_{fe,aa}$, $\exp(\cdot)$ the exponential map (Richard M. Murray & Sastry 1993), and $e_i \in \Re^3$ the unit basis vector; and $p^h_{h,fe}$, $p^{fe}_{fe,aa}$ and $p^{aa}_{aa,mp}$ are the anatomical lengths, which are assumed constant and known with $p^{fe}_{fe,aa} = de_1 = [d; 0; 0]$ (i.e., offset d along the PS motion axis - see Fig. 3.3. Here, with the IMU sensors attached to the $\{h\}$-frame and the $\{mp\}$-frame (see Fig. 3.2), we can directly measure $R^s_{s,h}$ and $R^s_{s,mp}$, and, consequently, $R^h_{h,mp}(\theta_{fe}, \theta_{ps}, \theta_{aa})$ from (3.2). By solving the inverse kinematics for $R^h_{h,mp}(\theta_{fe}, \theta_{ps}, \theta_{aa})$ with θ_{fe}, θ_{ps} and θ_{aa} being the pitch, roll and yaw angles, we can decode $(\theta_{fe}, \theta_{ps}, \theta_{aa})$ from $R^h_{h,mp}$ (Richard M. Murray & Sastry 1993), which are then in turn used to compute the full thumb posture. In this thesis, $(p^h_{h,fe}, p^{fe}_{fe,aa}, p^{aa}_{aa,mp})$ (and similar length parameters) are also off-line tuned to produce graphically-plausible motion during all of our experiments - how to on-line calibrate them is a research topic by itself and a topic of our future research as well.

On the other hand, to describe the posture of the index and middle fingers, we attach the $\{fe\}$-frame and $\{aa\}$-frame to the two-DOF MCP joint and the $\{pp\}$-frame to the proximal phalange as shown in Fig. 3.3. Then, similar to (3.1) with $R^{fe}_{fe,aa} = I$ and $d = 0$ (i.e., $\{fe\}$-frame is the same as $\{aa\}$-frame for the MCP joint), We can obtain the following kinematics similar:

$$\bar{g}^s_{s,pp} = \bar{g}^s_{s,h} \cdot \bar{g}^h_{h,pp}(R^h_{h,pp}, p^h_{h,pp})$$

where

$$\begin{aligned} R^h_{h,pp} &= R^h_{h,fe} R^{fe}_{fe,pp} = R^{s,T}_{s,h} R^s_{s,pp} \\ p^h_{h,pp} &= p^h_{h,fe} + R^h_{h,fe} p^{fe}_{fe,pp} \end{aligned}$$

where $R^s_{s,h}$ and $R^s_{s,pp}$ are measured by the IMU sensors attached respectively to the $\{h\}$-frame and the $\{pp\}$-frame, $R^h_{h,fe} = \exp(\theta_{fe} e_2)$ and $R^{fe}_{fe,pp} = \exp(\theta_{aa} e_3)$ with $(\theta_{fe}, \theta_{aa})$ decodable via the inverse kinematics of $R^h_{h,pp}(\theta_{fe}, \theta_{aa})$ similar for the thumb motion. Finally, for the MCP joint of the thumb or PIP joint of the index/middle fingers, one soft sensor is attached to provide the angle measurement θ_i or θ_j (along the e_2-direction). We can then obtain $\bar{g}^{mp}_{mp,pp}$ or $\bar{g}^{pp}_{pp,ip}$ with $R^{mp}_{mp,pp} = \exp(\theta_i e_2)$ or $R^{pp}_{pp,ip} = \exp(\theta_j e_2)$ and $p^{mp}_{mp,pp}$ or $p^{pp}_{pp,ip}$, with which we can complete the posture estimation of the thumb/index/middle fingers.

3.2.3. *Calibration method*

Each person has different size/shape of the finger/hand. Thus, the sensor attachments would be all different among different users, even if they wear the same FTM. To solve this issue, we perform a known-pose-based sensor calibration. First, the soft sensor has a linear relationship between the relative joint angle and its measurement $\rho_i \in \Re$, i.e.,

$$\theta_i = \beta_0 + \beta_1 \rho_i$$

where $\beta_0, \beta_1 \in \Re$ are coefficients. To find these coefficients, we need to take at least two known poses while measuring ρ_i with corresponding θ_i (e.g., ρ_i at $0°$ and $90°$). On the other hand, the IMU sensors provide orientation information expressed in the $\{s\}$-frame. However, whenever attached to the FTM and worn by the user, their real attachment is unknown and, in general, not the same as the target finger/hand segment as shown in Sec. 3.2.2. In other words, for each IMU, we have the following relation:

$$R^s_{s,f} = R^s_{s,b} R^b_{b,f}$$

where $R^s_{s,f}, R^s_{s,b}, R^b_{b,f} \in \mathrm{SO}(3)$ are the rotation of the finger/hand segment expressed in the $\{s\}$-frame, the measurement of the IMU, and the misalignment between the finger segment and the IMU sensor frame, respectively. Then, by letting user to assume an known pose, $R^s_{s,f}$ is known, $R^s_{s,b}$ is measured, thus, we can estimate $R^b_{b,f}$. Now, note that the number of unknowns is three (for IMUs) and two (for soft sensors). Thus, if we ask the user to assume three known poses, we can calibrate those three (or two) unknown quantities for each sensor. This is captured by the three calibration procedure with the three known postures, that is, 1) Align the direction of the tip of each index/middle finger and hand while keep them straight (i.e., $\theta_i = 0°$). 2) Align the direction of the tip of the thumb to be parallel to the direction of the step 1 while keep it straight. 3) Bend each PIP joint of the index/middle fingers and MCP joint of the thumb to be $90°$.

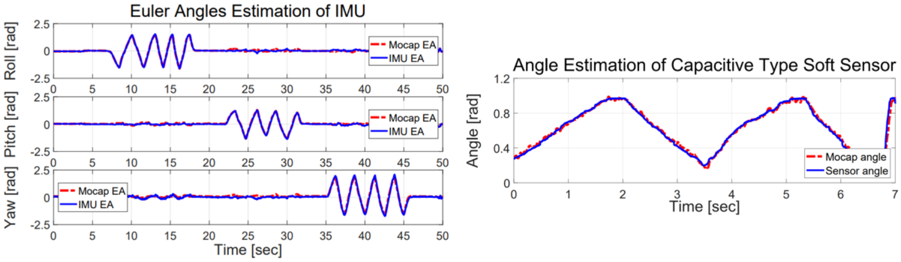


FIGURE 3.4. Comparison of ZYX Euler angles (EA) of the IMU (left) and single-DOF joint angle tracking of the soft sensor (right). Reference Euler angles are captured from MOCAP and palm-shaped (IMU) and index-finger-shaped (soft sensor) 3D printed mock-up. The mean angle errors are 1.567° and 0.0685° respectively.

3.3. Evaluation for VR Haptic Interaction Task

3.3.1. *Quantitative evaluation of FTM*

To precisely evaluate the performance of the FTM, we make 3D printed mock-ups while employing the MOCAP system (Optitrack) for the ground truth data acquisition. For the performance evaluation of the FTM, we first check the rotation estimation performance of the IMU. We rotate the IMU about 90° with respect to X/Y/Z-axis directions respectively to clearly check roll/pitch/yaw angle estimations, while attaching IR-markers for the MOCAP rotation tracking. In Fig. 3.4, we display the rotation tracking performance of the IMU expressed by the Euler angle. The mean Euler angle estimation error compared to the MOCAP is given as 1.567°. The small error may be originated from the imperfect calibration of the IMU (Markley 1988) and the latency of SO(3) filter (Mahony *et al.* 2008) which we use. On the other hand, we show the single-DOF joint angle tracking performance of the capacitive-type soft sensor employed, e.g., for the PIP joint of index finger. we achieves 0.0685° mean

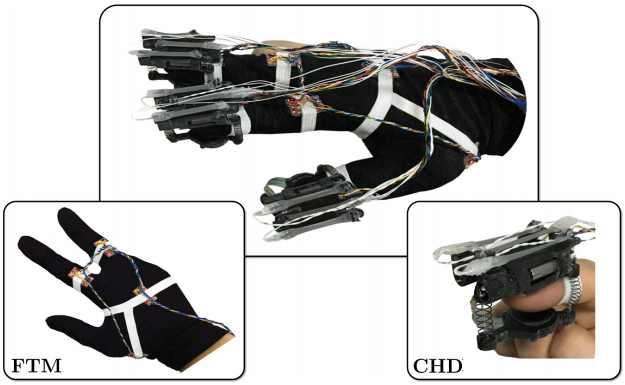


Figure 3.5. Wearable cutaneous haptic interface (WCHI) with finger tracking module (FTM) and cutaneous haptic device (CHD).

error where the small error mainly comes from the signal noise and the delay due to the 1st order low pass. Note that the smaller error than IMU may that we assume well-calibrated scenario of soft sensor (i.e., 3D-printed mock-up motion and apply it for only single-DOF motion). However, these accurate performance would imply that the rotational accuracy of FTM fair enough for finger-based interaction verified again by the previous results (chapter 2). While propagating these errors with the middle finger length in (Peters *et al.* 2002) (i.e., total 10.5 [cm]), the finger-tip position error would be 0.15 [cm]. and less than 3.64 [cm] which is the indistinguishable threshold under haptic feedback in VR as proven earlier.

3.3.2. *Implementation of Wearable Cutaneous Haptic Interface*

So far, we introduce the hardware configuration and its estimation of FTM. Then, the integrated WCHI for multi-fingered haptic interaction is presented by attaching the CHD to the finger-tip of the glove of the FTM since both modules do not interfere functionally and mechanically with

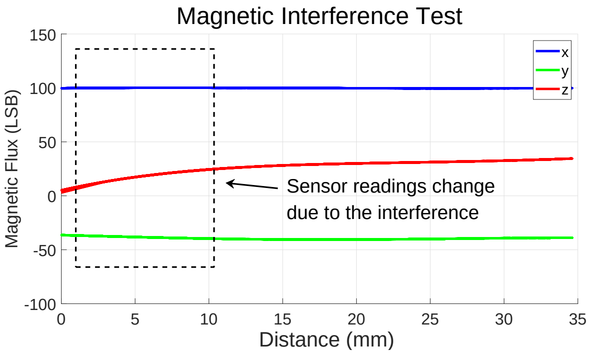


FIGURE 3.6. By using 3D-printed mock-up and changing the relative distance and rotation between IMUs and DC-micromotors, magnetic interference is evaluated by observing sensor measurements. Sensor measurements changes when the distance between the motor and the IMU becomes closer. However, it is observed that the interference is only significant if the distance is less than one centimeter.

each other. Here, the functional interference (i.e., electromagnetic interference between IMUs and DC-micromotors) is validated experimentally as shown in Fig. 3.6. For this, we make a 3D-printed mock-up where the motor is fixed with different orientations (i.e., 0°, 45°, and 90°) and the MEMS IMU moves along a fixed trajectory. We measure the relative distance with MOCAP and the IMU's magnetic flux readings, and educe the safe distance to be above 10 [mm] which is incorporated on design of each module and the integration for WCHI. Note that this safe distance is guaranteed even during full folding the fingers since the CHMs, thus DC-micromotors, are on the finger-tips while IMUs are on the proximal phalanges.

We also utilize two MCU boards (e.g., Arduino Nano) for the data acquisition of each IMUs and soft sensors of the FTM, which run at 200 [Hz] and 1 [kHz] respectively. One MCU board (i.e., Arduino Uno with the Adafruit Motor Shield V2 and standard op-amp circuit for sensor signal) is also employed for the control and the data acquisition of the CHM, which runs at 120 [Hz].

3.3.3. *Usability evaluation for VR peg-in-hole task*

Now, we conduct the user study to assess the effectiveness of the WCHI for VR application. We emulate a virtual manipulation task, inserting a breakable peg into a horizontally placed hole as shown in Fig. 3.7. The peg is 186.2 [mm] in height, 25.84 [mm] in radius, and 500 [g] in weight, which models a round bottle as a daily object. This peg inserting task is chosen here since, to manipulate its attitude and do the task, more complex finger control is required. All the subjects also attempt finger-tip manipulation, instead of power grasping, since it is fairly difficult to properly control the motion and insertion force without no haptic feedback on the palm. Our hypothesis is that the AA motion will be more important for this kind of real-life like complex task as compared to, e.g., the needle insertion (Meli *et al.* 2013) or the delivery task of simple object such as an egg, since the peg attitude should be controlled precisely to be inserted. On the other hand, we set the peg to be broken with large contact force (i.e., $\geqslant$ 5N). Therefore, subjects have to utilize the haptic feedback, even though it only exists at finger-tips, to successfully perform the task. The virtual hand is then controlled via the virtual coupling technique (Kim *et al.* 2017) where the desired hand motion is obtained from the FTM. The three-DOF contact force between the virtual hand and peg is fed back through CHD. Consequently, the test setup consists of WCHI, Oculus Rift HMD, soundproof earmuffs, and two HTC VIVE trackers to locate global position of a wrist and HMD respectively in a designated

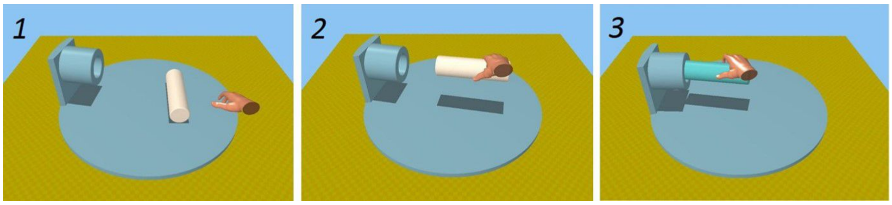


FIGURE 3.7. For each trial of the virtual peg-in-hole task, we randomly change both initial starting point of the peg and location of the hole (1). Then, a subject picks up and manipulates the peg to do the insertion task (2). The subject can see whether the peg is broken or not with the change of its color (3).

space as shown in Fig.3.1.

We design four different test settings to evaluate the performance of WCHI, especially the importance of the three-DOF cutaneous haptic feedback and the AA tracking motion of a hand for the VR application since typical VR hand interfaces have at best single-DOF haptic feedback and/or hand tracking without AA motion (e.g., MANUS VR glove, etc.). We intentionally turn on and off the actuation for with and without cutaneous haptic feedback (wHF or woHF) from CHM, and also turn on and off the allowance of AA tracking motion (wAA or woAA) of the FTM during the tests. As a result, the four settings are: 1) wHF wAA, 2) wHF woAA, 3) woHF wAA, and 4) woHF woAA. We then measure the task completion time for each trial and consider it as a performance measurement of the given task.

Then, the user study procedure consists of three phases: 1) familiarization, 2) main task, and 3) subjective questionnaire. During the familiarization, we introduced WCHI to users and verbally explain overall information about the task. We informed that there were four different settings, yet we did not provide details of each setting, not to make presuppose superiority and/or inferiority of each setting and try to distinguish them intentionally. Here, we also calibrate the WCHI, especially the FTM, to fit each subject's hand motion as in Sec. 3.2.

After the explanation and calibration, the 6 minutes of familiarization phase consists of two scenarios to gradually learn about the WCHI and HMD worn VR environment. For the first 3 minutes, users were instructed to touch and grasp the peg, which is suspended in the air by spring, and feel corresponding haptic feedback. We set the breaking force threshold of peg to be 5 [N] (i.e., $0 \leqslant |\lambda_N| \leqslant 10$ [N]) and change the peg color from white to blue (See (3) of Fig. 3.7). Users can learn an appropriate grasping force by matching the visual information and haptic feedback. In the next 3 minutes, users were asked to gently lift and rotate the peg from the ground to become accustomed to the peg manipulation with different hand postures. Throughout this phase, we provided users the full haptic feedback and hand tracking, i.e., setting 1 (wHF and wAA).

After then, each user experienced total 20 main tasks with repeated 4 different randomized settings for 5 times to minimize the learning effect. We also randomly (yet not too much) change the hole position and the starting point of the peg, again to minimize the learning effect. During the task, once the peg is broken down, it has to move back to a starting point while task time is continuously running. One thing to mention here is that we considered first 4 tasks as an extension of familiarization phase for the main task. Thus, we took account the results of last 16 tasks as a valid data for the analysis.

Ten users participated in the study including 9 male and 1 female in average 25.3±2.0 years old. All of them were right-hand except two users. All users did not have any physical or mental difficulty to complete the

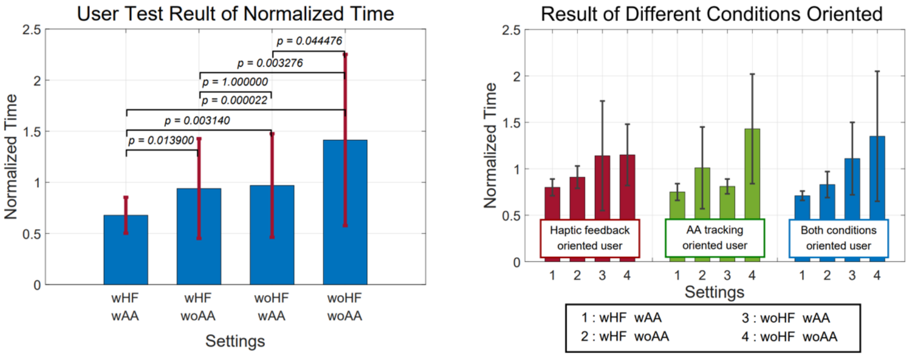


FIGURE 3.8. Normalized mean time and standard deviation results of ten users according to 4 different settings (left) and haptic feedback oriented user (user 5, red/right), AA tracking motion oriented user (user 4, green/right), and both condition oriented user (user 9, blue/right). There exists statistically significant difference between each setting except between setting 2 and setting 3 using the Bonferroni method ($p = 1.000$).

given test.

To evaluate the difference between each setting, we collected the total 160 trials data from ten users. The normalized mean time and standard deviation of ten users for each setting are depicted in Fig.3.8. As a result, the normalized mean time of setting 1 (0.678) takes about 2.1 times less than that of setting 4 (1.414). Also, the standard deviation of setting 1 (0.176) is 4.8 times less than that of setting 4 (0.838). On the other hand, the normalized mean time and the standard deviation of setting 2 (0.939 ± 0.488) and setting 3 (0.969 ± 0.507), where neither one of the conditions was not allowed, do not show significant difference. Both setting 2 and setting 3 take less time than setting 4 while taking more time than setting 1.

In further analysis, the normalized time is tested with one-way repeated measures ANOVA with the Greenhouse-Geisser correction ($\epsilon = 0.672$). The analysis determines that the normalized time has statistically significant difference between four settings on the peg-in-hole task, ($F(2.018, 78.684) = 13.076, p = 0.000012$). Post hoc tests using the Bonferroni method revealed that there exists statistically significant difference between each setting except between setting 2 and setting 3 as shown in Fig.3.8. This result does not change when we employ the Holm-Bonferroni method, which is known to be less conservative than the Bonferroni method. This is because, for our post hoc test, the p-value between setting 2 and setting 3 is one (i.e., $p = 1.000$). We illustrate every p-values between four settings in Fig. 3.8.

3.4. Discussion

We develop a novel wearable FTM, which is compatible with haptic devices. The successful integration with CHDs is presented and the efficacy of finger tracking and finger-tip haptic feedback accurately generated thanks to our FTM is also studied through the user test for dexterous VR task. Through the experimental results and we can deduce some insight for VR manipulation. First of all, on VR environment, each user uses and is affected by different conditions (or information). Three users are more likely to be affected by the role of haptic feedback than AA tracking motion. We called them haptic feedback oriented users. These users tend to depend more on haptic feedback on or off conditions than existence of AA tracking motion to complete the trials. Conversely, two users were affected by the role of AA tracking motion than haptic feedback, called them AA tracking oriented users. These users depend more on allowance of the AA tracking motion on VR manipulation than haptic feedback.

Rest users do affect by both conditions similarly. Above all, all ten users still performed consistently best with setting 1 where both haptic feedback and AA motion are provided. This result clearly shows that both haptic feedback and AA tracking motion are important for the virtual manipulation task.

Second, when we compare the haptic feedback and the AA motion, the difference in p-value size can be interpreted that the haptic feedback is more likely to be effective than the AA motion. For example, the p-value between setting 3 and setting 4 is 0.044476 which is larger and similar to the significance level while the p-value between setting 2 and setting 4 is 0.003276. This tendency is also found when we compare setting 2 and setting 3 to setting 1. This interpretation is consistent with the research context that many studies have focused more on the haptic feedback than on the AA motion, and can be an explanation of why the AA motion gets less attention for a hand interface while the haptic feedback is often pointed out for constructing immersive and informative VR interaction.

Chapter 4

Visual-Inertial Skeleton Tracking for Human Hand

4.1. Motivation

The wearable FTM in the previous chapter can fairly provide accurate tracking result (multi-DOFs motion (including AA motion) by IMU/compass and single-DOF motion by soft sensor) compatible with haptic devices. However, the proposed FTM cannot overcome the limitations of existing methods fundamentally. This is because the developed module is truly not fusing the complementary sensors, but only deploys the heterogeneous sensors for specific purpose (i.e., multi-fingered haptic interaction). Thus, the issues of each sensor remain unsolved, which means facing any issue of each sensing modality would lead to the unstable performance of the FTM For example, every problem oriented from compass (e.g., interaction with electronics), sensor drift, bias, different hand shape of users, dependency to external positioning sensor are still existing in the system, this is the motivation of developing a novel hand tracking framework, which overcome every issue of existing methods (occlusion (vision sensor), magnetic interference (IMU/compass), and mechanical contacts (soft sensor)), by developing complete sensor fusion framework.

To solve these issues, we propose a novel sensor fusion algorithm which would not require problematic compass or soft sensor, and rather utilize visual sensor, which has complementary aspects with IMU. Thanks to fusing visual information into inertial information, many parameters/disturbance can be eliminated from the system, or estimated precisely in real-time. In other words, we propose a novel **visual-inertial skeleton tracking (VIST)** system and its algorithm for accurate, robust and affordable hand tracking, while overcoming all the fundamental limitations of the other methods.

More specifically, we construct a sensor glove with seven IMUs and thirty-seven visual markers (of four different colors) attached on that and also with a head-mounted stereo camera; and a filtering-based visual-inertial hand tracking algorithm with hand anatomical constraints taken into account and also with auto-calibration of hand/sensor-related parameters. The stereo camera is used here merely owing to its availability and compatibility with VR/AR headsets; other vision sensors are equally applicable to our proposed VIST framework as explained in Sec. 4.2.3.

One of the key innovations of our VIST framework is that we fuse the visual and inertial sensors in a tightly-coupled (TC) manner (i.e., not only from visual to inertial (e.g., IMU drift correction (Tao *et al.* 2007; Bleser *et al.* 2011)), but also from inertial to visual (e.g., IMU-aided correspondence search in chapter 4.3). This TC-fusion is crucial to cope with the peculiarity of hand tracking, that is, a number of skeletons (i.e., fingers) are moving fast with occlusions among them in a small-size space (i.e., on the palm). This then necessitates us to utilize passive markers (for implementation/cost affordability), which are anonymous (up to different colors), as the space is too small to accommodate tagged visual markers (e.g., AR markers (Mallat *et al.* 2020), VIVE trackers (Li *et al.* 2019)), and whose number should be as many as possible for robustness against the occlusions. With these many anonymous visual markers, their correspondence search problem, central to the accurate vision processing, becomes very challenging, and, if it were not for this TC-fusion, our VIST algorithm simply fails the correspondence search with the tracking becoming unstable (See chapter 4.8).

In this chapter, we present detailed description about every component of hardware and algorithm of the proposed VIST framework. Then, the thorough evaluation for the developed VIST is performed in quantitative and qualitative manners. The experiments are designed to cover every possible scenarios for universal hand tracking system, that is, across from the case of normal free hand motion to the challenging scenario (visually complex background, object-interaction, wearing CHDs, and outdoor environment). The comparative studies with other existing methods are also presented to clearly show the superiority of our VIST.

4.2. Hardware Setup and Hand Models

4.2.1. *Human Hand Model*

The human hand is modeled as a segment-joint skeleton model (Wong *et al.* 2015) as shown in Fig. 4.1-A, where the types of the joints are determined according to their anatomical structures. This segment-joint model is widely adopted in many model-based hand tracking systems with vision, IMU/compass or soft sensors (Baldi *et al.* 2017; Lee *et al.* 2019; Mueller *et al.* 2017; Tkach *et al.* 2017). In this thesis, we choose the target tracking segments to be the dorsum of the hand and the three (thumb, index and middle) fingers, which play key roles in our daily activities and influence the motions of the ring/little fingers (Santello *et al.* 1998). We also assume the musculoskeletal dependencies (i.e., synergy (Hrabia *et al.* 2013; Baldi *et al.* 2017; Lee *et al.* 2019)) to estimate the angle of the interphalangeal (IP) joint from the metacarpophalangeal (MCP) joint for the thumb, and that of the distal interphalangeal (DIP) joints from the proximal interphalangeal (PIP) joints for the index and middle fingers. Since 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 can be straightforwardly done.

4.2.2. *Wearable Sensor Glove*

We fabricate a sensor glove based on our previous work (Lee *et al.* 2019), comprising of the 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 thirty-seven visual markers (fabric blobs with four different colors (red, blue, green and yellow). The sensor configuration is

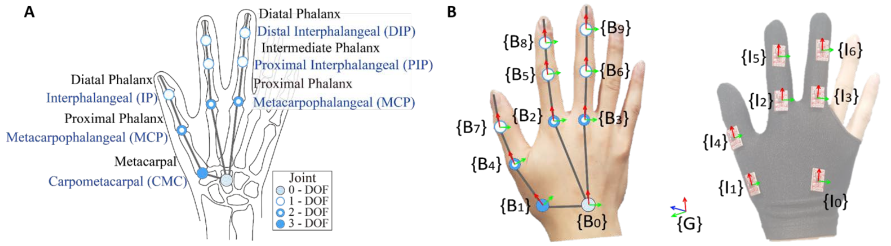


FIGURE 4.1. Schematic view of the wearable sensor glove: **(A)** Adopted anatomical model of the human hand and **(B)** Assignment of the coordinate frames and attachment of the IMUs.

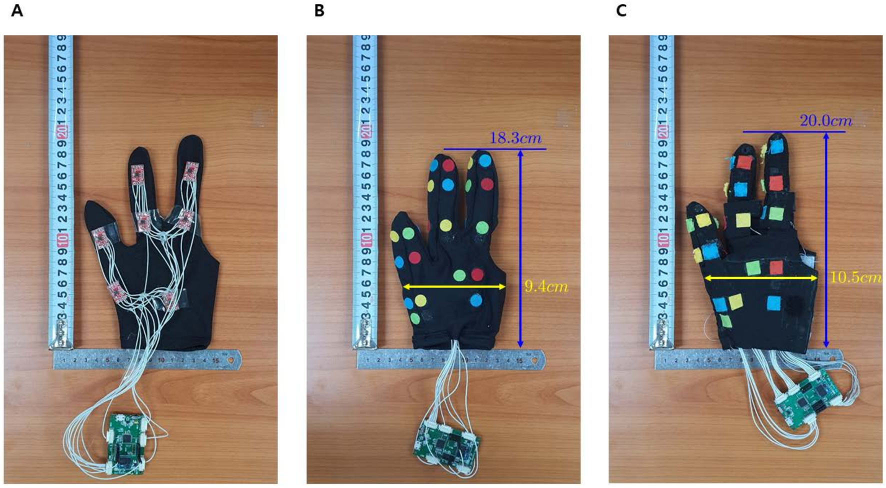


FIGURE 4.2. **(A)** A snapshot of the inner glove on which seven IMUs are attached. The sensor glove is fabricated by covering this inner glove with the outer glove where the visual markers are attached. **(B)** A snapshot of the smaller sensor glove attached with circular visual markers. Their length from the wrist to the middle finger-tip and the width from the thumb MCP to the right side of the hand are described in the figure. **(C)** A snapshot of the bigger sensor glove attached with square visual markers. Their length and the width are also described in the main text.

slightly modified from the previous one such that seven IMUs are attached as shown in Fig. 4.2.

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 if its corresponding hand-segment index i whenever relevant - see Fig. 4.1-B.

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). One the other hand, each $\{I_j\}$ is attached to its IMUs, whose pose (i.e., position and orientation) is not necessarily matched with corresponding $\{B_j\}$. Thus, many IMU-based tracking systems attempt to align $\{I_i\}$ with $\{B_j\}$ when attaching IMUs or require calibration before the operation through a sequence of indicated postures (Yuan *et al.* 2013; Luinge *et al.* 2007; Seel *et al.* 2014). However, misalignment error is inevitable when attaching IMUs, while the calibration is often not precise as there is always some human error to take the indicated postures. In contrast to that, our VIST algorithm contains the auto-calibration of such errors in real-time (see chapter 4.5), thereby, significantly improves tracking performance.

A user is recommended to wear a sensor glove slightly smaller than their hands to avoid sensor slippage on the hand from wearing larger gloves. The glove itself is 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 construct two sensor gloves of different sizes (Fig. 4.2). The length from the wrist to the middle finger's tip and width from the thumb MCP joint to the right side of hand are 18.3 cm and 9.4 cm for the smaller glove and 20.0 cm and 10.5 cm for the larger one. The total weight of the sensor gloves, including the markers, IMUs, and MCUs are only 52 g and 55 g, respectively.

We deploy low-cost commercial IMUs, MPU9250 (Invensense®), and connect them to a custom-built microcontroller unit (MCU) board based on ATmega328. This board collects the IMU data and sends it to the computer operating the main algorithm. The data acquisition from each IMU is at 100 Hz, and the data are sent through SPI communication protocol.

We adopt color blobs (circular/square color patches made from fabric) as the visual passive markers since they can be simply attached to the gloves without extra electronic installations (power, wire, or diode). Other types of anonymous markers are equally possible to the proposed VIST framework, according to the operating environment (e.g., gloves made with pattern-printed fabrics, IR/UV LEDs, reflective markers with IR cameras, or deep-learning-based features). The color blobs are attached to the designated positions of the sensor gloves (including positions directly above the IMUs), which are determined to ensure that appropriate numbers of markers could be seen from any viewpoint of the camera empirically. Fabric patch with four distinct hues (red, yellow, green, and blue) are employed in fabricating the markers. We attach 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.

4.2.3. *Stereo Camera*

We utilize Stereolabs® ZED Mini as the stereo camera, which is manufactured with affordable weight, size, and baseline to be comfortably equipped with a HMD: 62.9 g weight, 90° (H) x 60° (V) FOV, 63 mm baseline, and 1280x720 resolution for each image. Note that other types of vision sensors (e.g., monocular camera, depth/IR camera with IR markers) are also applicable for the proposed VIST framework by slightly modifying the marker detection/stereo matching processes, which are not core but replaceable module of the proposed algorithm. We assume the stereo camera mounted on the user's head part (e.g., equipped with a HMD, AR glass, or safety goggle), since this configuration makes our system fully portable without installing external sensors and comparable to recent commercial HMDs, which are normally equipped with two or more cameras (e.g., HTC vive pro, windows MR, Hololens).

4.3. Visual Information Extraction

Using the mentioned models, we design our VIST algorithm to be composed of the following two parts (Fig. 4.3-A): **visual information extraction** and **visual-inertial hand motion estimation**. The visual information extraction process comprises three sub-processes (Fig. 4.3-B): marker detection in raw images, left-right stereo matching, and IMU-aided correspondence search, through which the 3D positional observations of the visual markers are robustly matched to the actual markers on the glove.

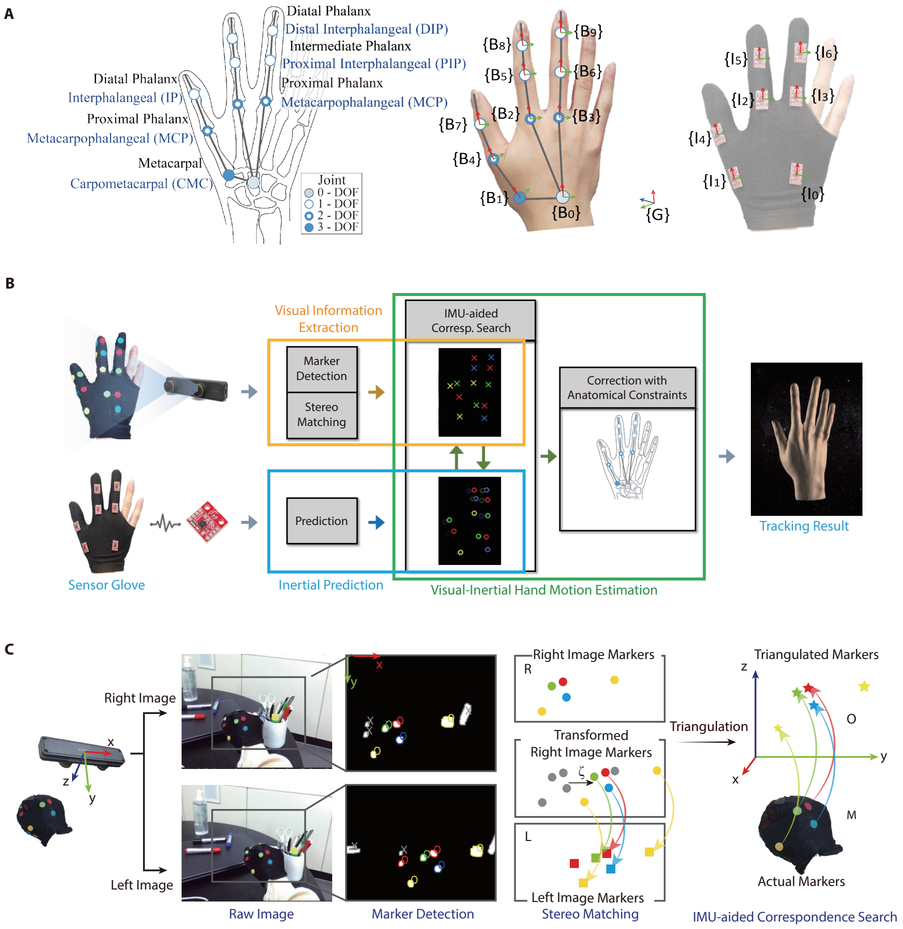


FIGURE 4.3. **(A)** Overview of the proposed visual-inertial skeleton tracking (VIST) algorithm: IMU-predicted information (left-down) are fused with the visual information (left-up). **(B)** Pipeline of the visual information extraction: (1) marker detection: only blobs satisfying color and shape requirements are recognized as 2D observations, $\mathcal{R}$ and $\mathcal{L}$. (2) left–right stereo matching: $\mathcal{R}$ and $\mathcal{L}$ are triangulated as 3D positional observations $\mathcal{O}$. (3) IMU-aided correspondence search: $\mathcal{O}$ are assigned to actual markers on the glove $\mathcal{M}$ based on observation probabilities.

4.3.1. *Marker Detection in Raw Images*

To detect the visual markers (i.e., color blobs) in the raw stereo images, we utilize the following two requirements standard in the field of computer vision: 1) hue-saturation-values (HSVs) requirement, that is, we extract only the visual blobs having HSV within the predefined intervals of the blob colors; 2) 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., $\mathcal{R} = \{r_1, r_2, ...r_{N_R}\} \in \Re^{2N_R}$ for the right image and $\mathcal{L} = \{l_1, l_2, ...l_{N_L}\} \in \Re^{2N_L}$ for the left image, respectively.

4.3.2. *Cost Function for Point Matching*

To fuse visual-inertial sensor data, two sequential sub-processes, that is the stereo matching and the correspondence search which will be described in the following sections, should be solved. We formulate two sub-processes as point set registration problem, which finds the transformation parameter θ (translation, rotation, scale, etc.) to overlap a scene point set $\mathcal{S} = \{s_1, s_2, ...s_N\}$ with a model point set $\mathcal{Y} = \{y_1, y_2, ...y_M\}$. Among the point set registration algorithms, we deploy coherent point drift algorithm (CPD) introduced in (*55*), which represents the model set $\mathcal{Y}$ as the Gaussian mixture model (GMM) and finds the transformation parameter by maximizing the GMM posterior probability, which is given by

$$p(\mathcal{S}|\mathcal{Y}, \theta) = \prod_{i=1}^{N} p(s_i|\mathcal{Y}, \theta) \tag{4.1}$$

Given the model set $\mathcal{Y}$, the observation probability of a scene point s_i is determined by the product of the probability where s_i corresponds to each model point y_j or outlier, which is given by

$$p(s_i|\mathcal{Y}, \theta) = (1 - w) \sum_{j=1}^{M} p(y_j) p(s_i|y_j, \theta) + w/N \tag{4.2}$$

$$p(s_i|y_j, \theta) \sim \mathcal{N}(T(y_j, \theta), \Sigma) \tag{4.3}$$

where $T(y_j, \theta)$ is the transformation of the point y_j using the parameter θ, w is the parameter determining outlier ratio, Σ is the covariance matrix of the scene points and $p(y_j)$ is the prior probability which is assumed to be even probability $1/M$.

4.3.3. *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 utilize coherent point drift (CPD) algorithm (Myronenko & Song 2010), which is a classical point set registration method. More specifically, to match 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 point sets) using Expectation Maximization (EM) algorithm to maximize the product of the correspondence probability and the transformed GMM probability of the other point set. We choose the CPD algorithm in this thesis due to its simplicity in the rigid point set registration. Other methods (e.g., (Hirose 2020; Gao & Tedrake 2019)) may also be used depending on the matching complexity.

In the stereo matching, without loss of generality, we represent the observation set of the right image $\mathcal{R}$ as a GMM, and define a transformation parameter between $\mathcal{L}$ and $\mathcal{R}$ as $\zeta \in \Re^1$ to represent the 1-DOF horizontal parallax between the right/left point sets and neglecting vertical transformation as the raw stereo images are assumed to be rectified. Given the right observation set $\mathcal{R}$ and transformation ζ, the GMM likelihood function of the left observation set $\mathcal{L}$ is defined as

$$\mathrm{p}(\mathcal{L}|\mathcal{R}, \zeta) = \prod_{h=1}^{4} \prod_{i=1}^{N_{L,h}} \mathrm{p}(l_{i,h}|\mathcal{R}_h, \zeta) \tag{4.4}$$

where $h = \{1, 2, 3, 4\}$ are the indexes of the four hues, and $N_{L,h}$ is the number of left observed markers for each hue h. Note that we divide the 2D marker observation sets into four groups based on the color hues (i.e., red, blue, green, and yellow).

Since $\mathcal{R}$ is the model point set represented by the GMM centroids, the probability of each left observation $l_{i,h} \in \mathcal{L}$ is given as

$$\mathrm{p}(l_{i,h}|\mathcal{R}_h, \zeta) = (1 - w_s) \sum_{j \in \mathcal{R}_h} \mathrm{p}(r_j)\mathrm{p}(l_{i,h}|r_j, \zeta) + w_s/N_{L,h} \tag{4.5}$$

$$\mathrm{p}(l_{i,h}|r_j, \zeta) \sim \mathcal{N}(T(r_j, \zeta), \Sigma_s) \tag{4.6}$$

where $T(r_j, \zeta)$ is the transformation of the right point r_j using parameter ζ, w_s is the parameter determining the outlier ratio of stereo matching, Σ_s is the covariance matrix of pixel observation noises of the adopted camera (tuned by pilot tests), and $\mathrm{p}(r_j)$ is the prior probability assumed to be the even probability of $1/N_{R,h}$. The transformation parameter ζ is obtained by maximizing the likelihood function (4.4) using the EM algorithm.

The transformation parameter ζ is then obtained by using the EM algorithm as in (Myronenko & Song 2010). Once ζ is determined, each left point l_i has a matched candidate $r_{j,\min}$, which is the closest right point when transformed by ζ. As it is plausible that only one blob is observed from a (left or right) camera, we define two additional conditions to identify such blobs as outliers: distance from the closest point is less than a predefined threshold $||l_i - T(r_{j,\min}, \zeta)||^2 < \delta_1$; ratio of the first to second closest points $r_{j,\min2}$ is less than another predefined threshold $||(l_i - T(r_{j,\min}, \zeta))/(l_i - T(r_{j,\min2}, \zeta))||^2 < \delta_2$. When l_i and $r_{j,\min}$ satisfy the two conditions simultaneously, the points are matched; otherwise, l_i is identified as an outlier. Then, the matched points are triangulated using mechanical specifications (i.e., focal length and baseline) of the adopted camera, which constructs the positional observations of the markers $\mathcal{O} = \{o_1, o_2, \dots, o_{N_\mathcal{O}}\} \in \Re^3$, where $N_\mathcal{O}$ is the number of final matched points from the stereo images.

Once ζ 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 ζ. As it is plausible that only one blob is observed from one (left or right) camera, we define two additional conditions based on Euclidean distance to identify such a blob as outliers. Mathematical equations related to the stereo matching are described as follows. When l_i and $r_{j,\min}$ satisfy the 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, $\mathcal{O} = \{o_1, o_2, \dots, o_{N_O}\} \in \Re^3$, are constructed, where N_O is the number of the matched points from the stereo images.

4.4. 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 EKF propagation in following chapter), $\mathcal{M} = \{m_1, m_2, ...m_{N_M}\} \in \Re^3$, where $N_M = 37$ (i.e., the number of all the visual markers attached to the glove). We then again apply the CPD algorithm to match $\mathcal{O}$ and $\mathcal{M}$, by defining $\mathcal{M}$ as a GMM and finding the transformation parameter $\eta \in \Re^3$ between $\mathcal{O}$ and $\mathcal{M}$, which is the 3-DOF translation between the two sets assuming that the rotations of the predicted marker set $\mathcal{M}$ and the latest observation set $\mathcal{O}$ are aligned well, since the rotation of $\mathcal{M}$ can be updated fairly precisely with the gyroscope over a short period of time (Lee *et al.* 2019). The GMM likelihood function of the set $\mathcal{O}$ is then defined as

$$\mathrm{p}(\mathcal{O}|\mathcal{M}, \eta) = \prod_{h=1}^{4} \prod_{i=1}^{N_{O,h}} \mathrm{p}(o_{i,h}|\mathcal{M}_h, \eta) \tag{4.7}$$

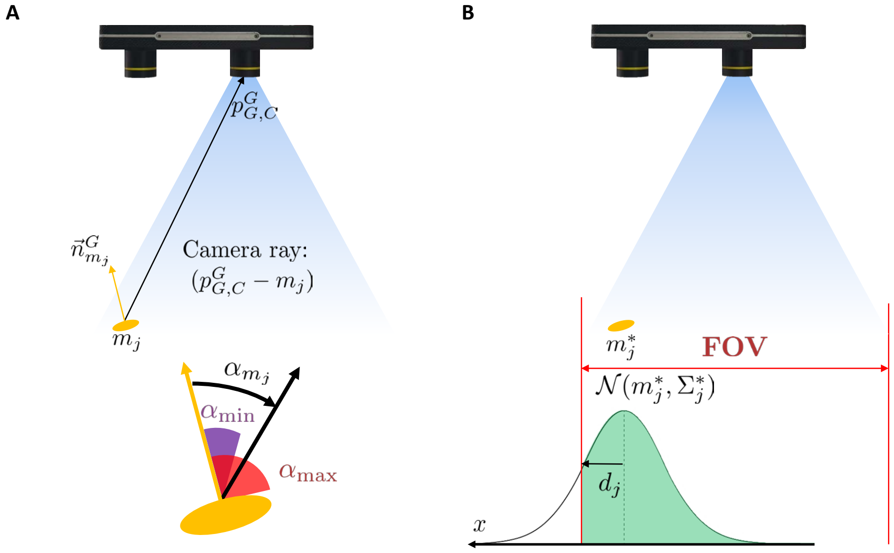
where $N_{O,h}$ is the number of marker observations for each hue $h \in \{1, 2, 3, 4\}$ (i.e., red, blue, green and yellow).

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

$$\mathrm{p}(o_{i,h}|\mathcal{M}_h, \eta) = (1 - w_c) \sum_{j \in \mathcal{M}_h} \mathrm{p}(m_j)\mathrm{p}(o_{i,h}|m_j, \eta) + w_c/N_{O,h} \tag{4.8}$$

$$\mathrm{p}(o_{i,h}|m_j, \eta) \sim \mathcal{N}(T(m_j, \eta), \Sigma_c) \tag{4.9}$$

where $T(m_j, \eta) \in \Re^3$ is the transformation of the point $m_j \in \mathcal{M}$ using the parameter $\eta \in \Re^3$, $w_c \in [0, 1]$ is the parameter determining the outlier ratio of the correspondence search, and $\Sigma_c \in \Re^{3\times3}$ is the covariance matrix of observation noises of the marker triangulation, which is obtained by pilot tests (Fig. 4.9).



$$\mathbf{p_{n,R}}(\mathbf{m_j}) = \begin{cases} 1, & (\alpha_{v_j} \leq \alpha_{\text{min}}) \\ \frac{\alpha_{\text{max}} - \alpha_{m_j}}{\alpha_{\text{max}} - \alpha_{\text{min}}}, & (\alpha_{\text{min}} < \alpha_{m_j} \leq \alpha_{\text{max}}) \\ 0, & (\alpha_{\text{max}} < \alpha_{m_j}) \end{cases}$$

$$\mathbf{p_{f,R}}(\mathbf{m_j}) = \int_{-\infty}^{d_j} (e^{-(x^2)/(2\sigma_j^{*2})})/(\sqrt{2\pi\sigma_j^{*2}})dx$$

FIGURE 4.4. Example figure depicting how to obtain the observation probability of a marker m_j from the right camera **(A)** The probability from camera-facing factor: when the normal vector of the marker $\vec{n}^{G}_{m_j}$ is computed by IMU-predicted hand pose, the probability $\mathrm{p}_{n,R}(m_j)$ is determined by the angle between this vector with the camera ray $p^{G}_{G,C} - m_j$. $\alpha_{\text{min}}, \alpha_{\text{max}}$ are defined according to the adopted type of the visual marker (e.g., $\alpha_{\text{min}} = 45°$ and $\alpha_{\text{max}} = 90°$ for our color blob marker through a pilot test). **(B)** The probability from FOV factor: when the 2D projected position m^*_j and the projected covariance σ^*_j of the visual marker m_j is predicted by IMU, the FOV probability $\mathrm{p}_{f,R}(m_j)$ is determined by accumulating the area inside the camera FOV of the probability distribution.

It is typical that only less than one third of all the (thirty seven) makers survive the stereo matching (i.e., average(N_O) = 10.44 - see Sec. 4.6. Further, since those markers in $\mathcal{O}$ are anonymous (up to different colors) and the 3D motion of each finger is precarious/fast, if we attempt the correspondence search only with the vision information via CPD (i.e., $\hat{\mathcal{O}}_{t-1} \rightarrow \mathcal{O}_t$, where $\hat{\mathcal{O}}_{t-1} \in \Re^{3\times37}$ is the computed positions of all the markers based on the hand motion estimated at the previous time $t-1$, while $\mathcal{O}_k \in \Re^{3N_O}$ 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/chattering (see Fig. 4.8-D). 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 then renders our VIST algorithm TC-fusion with its accuracy and robustness drastically improved. More precisely, we consider the following factors: camera-facing factor and FOV factor (Fig. 4.4).

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 $\mathrm{p}_n(m_j)$ for the camera-facing factor is defined as

$$\mathrm{p}_n(m_j) = \begin{cases} 1, & \text{for } \alpha_{m_j} \leqslant \alpha_{\min} \\ (\alpha_{\max} - \alpha_{m_j})/(\alpha_{\max} - \alpha_{\min}), & \text{for } \alpha_{\min} < \alpha_{m_j} \leqslant \alpha_{\max} \\ 0, & \text{for } \alpha_{\max} < \alpha_{m_j} \end{cases} \tag{4.10}$$

where ($\alpha_{\min}$, $\alpha_{\max}$) is the visible angular range of the adopted marker type, and α_{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 $\mathrm{p}_f(m_j)$ for the FOV factor is given by integrating the area inside the FOV of the Gaussian distribution of m_j^*, which is defined as

$$\mathrm{p}_f(m_j) = \int_{-\infty}^{d_j} (e^{-(x^2)/(2\sigma_j^{*2})})/(\sqrt{2\pi\sigma_j^{*2}})dx \tag{4.11}$$

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

Using the camera-facing factor (4.10) and the FOV factor (4.11), the final observation probability of the marker m_j is the product of these two probabilities for both cameras, that is,

$$\mathrm{p}(m_j) = \prod_{c=R,L} \mathrm{p}_{n,c}(m_j)\mathrm{p}_{f,c}(m_j) \tag{4.12}$$

where c is the index of both the right and left cameras. We also eliminate an marker m_j occluded from the other segments in the point matching (i.e., $\mathrm{p}(m_j)$ 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 $\mathrm{p}(m_j)$ is to compute the posterior probability (4.8) for the correspondence search.

Once the transformation parameter $\eta \in \Re^3$ for (4.7) 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 $\mathrm{p}(m_j|o_i)$ from all the markers $\mathcal{M} = \{m_1, m_2, ...m_{N_\mathcal{M}}\}$, that is defined as

$$\mathrm{p}(m_j|o_i) = \mathrm{p}(o_i|m_j)\mathrm{p}(m_j)/\mathrm{p}(o_i) = (\mathrm{p}(o_i|m_j)\mathrm{p}(m_j))/(\sum_{j=1}^{N_\mathcal{M}} \mathrm{p}(o_i|m_j)\mathrm{p}(m_j)) \tag{4.13}$$

An observation o_i is assigned to marker m_j with the maximum matching probability $\mathrm{p}(m_j|o_i)$ from among all markers $\mathcal{M} = \{m_1, m_2, ...m_{N_\mathcal{M}}\}$.

However, this corresponds to multiple observations $\{o_i, o_j, ...\}$ for the marker m_j. To reject pairs with duplicate matches as in stereo matching, we introduce thresholding condition based on Euclidean distance to identify outliers as follow. Only one pair with the minimum Euclidean distance is selected, and if its matching probability $\mathrm{p}(m_j|o_i)$ is less than the predefined threshold δ_3, o_i is identified as an outlier (i.e., false observation for the visual marker). Then, finally, we attain the IMU-aided correspondence search: $\mathcal{M} \supset \hat{\mathcal{Z}} \to \mathcal{Z} \subset \mathcal{O}$, where $\hat{\mathcal{Z}}$ and $\mathcal{Z}$ are the sets of the corresponded markers in $\mathcal{M}$ and $\mathcal{O}$ with the same dimension and will be used for the EKF update (4.26).

4.5. Filtering-based Visual-Inertial Sensor Fusion

For the visual-inertial sensor fusion with a large number of states at a rate faster than the sensor sampling rate, we deploy the extended Kalman filter (EKF), which is the most common estimator for nonlinear systems and exhibits reasonable performance with limited computation load (Table 4.2). This EKF consists three sub-processes: prediction with IMU information, correction with visual information, and correction with anatomical constraints.

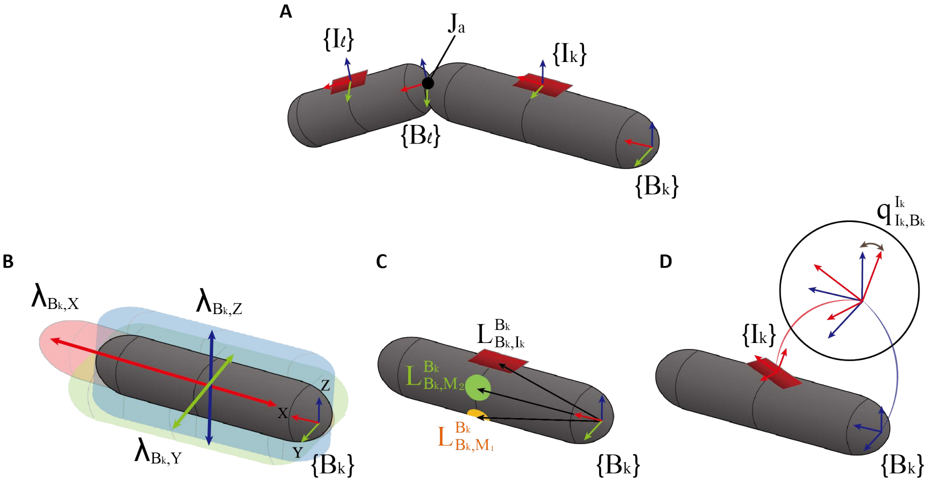


FIGURE 4.5. **(A)** The simplified figure depicting the k-th and l-th segments which are connected on the pivot joint J_a. **(B)** The scale parameter $\lambda_{B_k} := [\lambda_{B_k,X}; \lambda_{B_k,Y}; \lambda_{B_k,Z}] \in \Re^3$ represents the size of the corresponding segment with respect to the segment body frames $\{B_k\}$. **(C)** The vectors of two markers (m_1, m_2) and the IMU (I_k) relative to the attached segment $\{B_k\}$ are predesignated in the fabrication process. Note that these vectors are multiplied by estimated hand scale λ_{B_k} in the correction step for marker measurement as described in the main text. **(D)** The parameter for IMU attachment offset $q^{I_k}_{I_k,B_k} \in \Re^4$ is a unit quaternion, which represents the misalignment between the body coordinate frame $\{B_k\}$ and the IMU coordinate frame $\{I_k\}$.

4.5.1. *EKF States for Hand Tracking and Auto-Calibration*

We define the EKF states for each segment of the hand (total seven segments) as

$$x := [x_s; x_p] \in \Re^{23} \tag{4.14}$$

where $x_s := [p^G_{G,I}; v^G_{G,I}; q^G_{G,I}; b_g; b_a] \in \Re^{16}$ is the motion-related states 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 (Trawny & Roumeliotis 2005). On the other hand, we define the states for the auto-calibration of hand/sensor-related kinematic parameters s.t.,

$$x_p := [\lambda_B; q^I_{I,B}] \in \Re^7 \tag{4.15}$$

where $\lambda_B \in \Re^3$ is the scale factor of the attached segment $\{B\}$, which is dependent on the user hand size, and $q^I_{I,B} \in \Re^4$ is the quaternion for misalignment between $\{I\}$ and $\{B\}$, which is dependent on the user hand shape and may also changes for each fitting (Fig. 4.5).

Inclusion of this auto-calibration 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 the 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 of course not true in practice and results in typically less accurate calibration and finger-tip position tracking. In contrast, due to the real-time/auto-calibrations of x_p utilizing the visual-inertial fusion, our proposed VIST framework can substantially improve tracking accuracy and use convenience as compared to the other systems.

4.5.2. *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:

$$\dot{\hat{x}} = f(\hat{x}, a_m, w_m) \tag{4.16}$$

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.

More specifically, this kinematic model (4.19) is derived as follows. The motion-related state $x_s = [\ p^G_{G,I};\ \ v^G_{G,I};\ \ q^G_{G,I};\ \ b_g;\ \ b_a] \in \Re^{16}$ is predicted using inertial measurements from IMUs. We define the kinematic model of the true-state x_s adopting the renowned model (Trawny & Roumeliotis 2005), that is,

$$\begin{aligned}
\dot{p}^G_{G,I}(t) &= v^G_{G,I}(t) \\
\dot{v}^G_{G,I}(t) &= R^G_{G,I} a^I_{G,I}(t) + g^G \\
\dot{q}^G_{G,I}(t) &= 0.5\Omega(w^I_{G,I}(t)) q^G_{G,I}(t) \\
\dot{b}_g(t) &= n_{bg}(t) \\
\dot{b}_a(t) &= n_{ba}(t)
\end{aligned} \tag{4.17}$$

where g^G is the gravity vector expressed in the global reference frame $\{G\}$, $a^I_{G,I}(t)$ and $w^I_{G,I}(t)$ are the true acceleration and angular velocity of the sensor respectively at time t, $R^G_{G,I}$ is the rotation matrix converted from the quaternion $q^G_{G,I}$ and $\Omega(\xi)$ is the matrix for the product of a vector $\xi \in \Re^3$ with a quaternion. The sensor biases are assumed to be a slow-varying drifts with zero-mean, white Gaussian noise processes n_{bg},

n_{ba}, the values of which are stated in the data sheet of the deployed IMU model.

The relationships between the true inertial value $a^I_{G,I}(t)$, $w^I_{G,I}(t)$ and the measurements $a_m(t)$, $w_m(t)$ from the IMUs are modeled as

$$\begin{aligned} w_m(t) &= w^I_{G,I}(t) + b_g(t) + n_g(t) \\ a_m(t) &= {R^I_{I,G}}^T(a^I_{G,I}(t) - g^G) + b_a(t) + n_a(t) \end{aligned} \tag{4.18}$$

where the IMU noise n_g and n_a are zero-mean, white Gaussian noise processes, which are also described in the data sheet.

The kinematic model of nominal-state $\hat{x}_s$ is given by

$$\begin{aligned} \dot{\hat{p}}^G_{G,I}(t) &= \hat{v}^G_{G,I}(t) \\ \dot{\hat{v}}^G_{G,I}(t) &= \hat{R}^G_{G,I}\hat{a}^I_{G,I}(t) + g^G \\ \dot{\hat{q}}^G_{G,I}(t) &= 0.5\Omega(\hat{w}^I_{G,I}(t))\hat{q}^G_{G,I}(t) \\ \dot{\hat{b}}_g(t) &= 0_{3\times1} \\ \dot{\hat{b}}_a(t) &= 0_{3\times1} \end{aligned} \tag{4.19}$$

where the estimated angular velocity is given by $\hat{w}^I_{G,I}(t) = w_m(t) - \hat{b}_g(t)$ and the estimated linear acceleration is given by $\hat{a}^I_{G,I}(t) = a_m(t) - \hat{b}_a(t)$.

We adopt error-state representation to reduce the computational load and guarantee the minimal system similar in (Trawny & Roumeliotis 2005). The nominal-state and error-state are defined by

$$\begin{aligned} \tilde{x}_s &= [\ \tilde{p}^G_{G,I};\ \ \tilde{v}^G_{G,I};\ \ \delta\theta^G_{G,I};\ \ \tilde{b}_g;\ \ \tilde{b}_a\ ] \in \Re^{15} \\ \tilde{x}_p &= [\ \tilde{\lambda}_B;\ \ \delta\theta^I_{I,B}\ ] \in \Re^6 \end{aligned} \tag{4.20}$$

where all the error states employ the additive error model except $\delta\theta$ which is the multiplicative attitude error. The kinematic model of the error-state

$\tilde{x}_s$ is defined by

$$\begin{aligned}
\dot{\tilde{p}}^G_{G,I}(t) &= \tilde{v}^G_{G,I}(t) \\
\dot{\tilde{v}}^G_{G,I}(t) &= -\hat{R}^G_{G,I}[\hat{a}^I_{G,I}(t)\times]\tilde{\theta}^G_{G,I}(t) - \hat{R}^G_{G,I}\tilde{b}_a(t) - \hat{R}^G_{G,I}n_a(t) \\
\dot{\tilde{\theta}}^G_{G,I}(t) &= -[\hat{w}^I_{G,I}(t)\times]\tilde{\theta}^G_{G,I}(t) - \tilde{b}_g(t) - n_g(t) \\
\dot{\tilde{b}}_g(t) &= n_{bg}(t) \\
\dot{\tilde{b}}_a(t) &= n_{ba}(t)
\end{aligned} \tag{4.21}$$

where $[\xi\times]$ is the skew-symmetric matrix of the vector $\xi \in \Re^3$.

The kinematics of the nominal-state $\hat{x}_p$ and error-state $\tilde{x}_p$ relating to the parameter x_p are defined by

$$\begin{aligned}
\dot{\hat{\lambda}}_B(t) &= 0_{3\times1} \quad \dot{\hat{q}}^I_{I,B}(t) = 0_{4\times1} \\
\dot{\tilde{\lambda}}_B(t) &= n_\lambda(t) \quad \dot{\tilde{\theta}}^I_{I,B}(t) = n_q(t)
\end{aligned} \tag{4.22}$$

where $n_\lambda \in \Re^3$ and $n_q \in \Re^3$ are zero-mean, white Gaussian noise processes, assuming to have slow-varying property similar to the IMU biases.

The error-state propagation model is defined by linearizing the aforementioned error-state dynamic model, that is,

$$\dot{\tilde{x}} = \begin{bmatrix} \dot{\tilde{x}}_s \\ \dot{\tilde{x}}_p \end{bmatrix} = \begin{bmatrix} F_s & 0_{15\times6} \\ 0_{6\times15} & 0_6 \end{bmatrix} \tilde{x} + Gn \tag{4.23}$$

$$F_s = \begin{bmatrix} 0_3 & I_3 & 0_3 & 0_3 & 0_3 \\ 0_3 & 0_3 & -\hat{R}^G_{G,I}[\hat{a}^I_{G,I}(t)\times] & 0_3 & -\hat{R}^G_{G,I} \\ 0_3 & 0_3 & -[\hat{w}^I_{G,I}(t)\times] & -I_3 & 0_3 \\ 0_3 & 0_3 & 0_3 & 0_3 & 0_3 \\ 0_3 & 0_3 & 0_3 & 0_3 & 0_3 \end{bmatrix} \tag{4.24}$$

$$G = \begin{bmatrix} 0_3 & 0_3 & 0_{3\times12} \\ -\hat{R}^G_{G,I} & 0_3 & 0_{3\times12} \\ 0_3 & -I_3 & 0_{3\times12} \\ 0_{12\times3} & 0_{12\times3} & I_{12} \end{bmatrix} \tag{4.25}$$

where F_s is the error-state transition matrix corresponding to the IMU sensor state, G is the input noise matrix, $0_{n\times m}$ is a n×m zero matrix and 0_n and I_n is a n×n zero matrix and identity matrix respectively. The system noise n is the sum of noise vectors defined by $n = [n_a;\ n_g;\ n_{b_g};\ n_{b_a};\ n_\lambda;\ n_q;] \in \Re^{18}$. Since the IMU measurements are sampled at the IMU sampling rate, we discretize the continuous-time prediction model for the VIST implementation.

4.5.3. *Correction with Visual Information*

The IMU-predicted motion of each segment is in general inaccurate owing to the lack of compass information, sensor noise and uncalibrated parameters (i.e., $b_g, b_a, \lambda_B, q^I_{I,B}$). Thus, we correct this IMU-predicted hand segment motion and also the uncalibrated parameters using the correspondence-matched marker measurements $\mathcal{Z} \subset \mathcal{O}$. More specifically, we utilize the linearized error model of the measurement equation s.t.,

$$\tilde{z}_{m_j} \ = \ z_{m_j} - \hat{z}_{m_j} \ \simeq \ H_{m_j}\tilde{x} + n_z \tag{4.26}$$

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 the equation (4.9). 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 derived as follows. Among all the segments, only the k-th segment attaching the j-th marker is affected by the j-th marker measurement. For the simplicity of expression, we represent observation matrix with respect to the state of the k-th segment and abbreviate the segment index k.

While the marker measurements are obtained with respect to the camera coordinates $\{C\}$, the sensor state $x_s = [\ p^G_{G,I};\ \ v^G_{G,I};\ \ q^G_{G,I};\ \ b_g;\ \ b_a]$ is defined with respect to the global frame $\{G\}$. Thus, the measurement m_j of the j-th marker is converted to z_{m_j} with respect to the global coordinate frame as

$$z_{m_j} \ = \ p^G_{G,C} + R^G_{G,C} m_j \tag{4.27}$$

where $p^G_{G,C}$ and $R^G_{G,C}$ are the transition and rotation matrices of the camera from $\{G\}$. To obtain the camera poses $p^G_{G,C}$ and $R^G_{G,C}$, we assume that a modularized camera localization algorithm, such as SLAM, VINS, or built-in localization algorithms is utilized, which are standard techniques for recent stereo cameras or HMDs. Note that if a camera localization algorithm adoption is unavailable, the proposed algorithm still tracks hand motions relative to the camera by simply modifying the reference coordinate frame of the sensor state $x_s = [\ p^G_{G,I};\ \ v^G_{G,I};\ \ q^G_{G,I};\ \ b_g;\ \ b_a]$ from $\{G\}$ to $\{C\}$.

Then, the marker measurement equation is given as follows:

$$\hat{z}_{m_j} \ = \ \hat{p}^G_{G,I} + \hat{R}^G_{G,I} \hat{R}^I_{I,B} d(\hat{\lambda}_B) L^B_{I,m_j} \tag{4.28}$$

where $d(\xi)$ refers to the 3×3 diagonal matrix of a vector $\xi \in \Re^3$, and L^B_{I,m_j} is the designated position of the marker m_j when fabricating the sensor glove (Fig. 4.5). One assumption for this measurement is that the surface of the glove is stretched proportional to the hand scale λ_B, as described in the Materials and Methods.

Given the marker measurement equation above,

$$\tilde{z}_{m_j} = z_{m_j} - \hat{z}_{m_j} \tag{4.29}$$

The observation matrix H_{m_j} of the j-th marker is defined by the linearization of this measurement equation, that is,

$$H_{m_j} = [H_{\tilde{p}^G_{G,I}} \;\; 0_{3\times3} \;\; H_{\tilde{\theta}^G_{G,I}} \;\; 0_{3\times6} \;\; H_{\tilde{\theta}^I_{I,B}} \;\; H_{\tilde{\lambda}_B}] \tag{4.30}$$

$$\begin{aligned} H_{\tilde{p}^G_{G,I}} &= -I_{3\times3} \\ H_{\tilde{\theta}^G_{G,I}} &= \hat{R}^G_{G,I}[\hat{R}^I_{I,B} d(\hat{\lambda}_B) L^B_{I,M_j} \;\; \times] \\ H_{\tilde{\theta}^I_{I,B}} &= \hat{R}^G_{G,I}\hat{R}^I_{I,B}[d(\hat{\lambda}_B) L^B_{I,M_j} \;\; \times] \\ H_{\tilde{\lambda}_B} &= \hat{R}^G_{G,I}\hat{R}^I_{I,B} d(L^B_{I,M_j}) \end{aligned} \tag{4.31}$$

where H_χ are the Jacobian of the measurement equation with respect to the error-state χ and $[\xi\times]$ is the skew-symmetric matrix of the vector $\xi \in \Re^3$.

Moreover, for delay-free estimations even in the case of fast hand motions, (Weiss 2012), we employ a ring buffer to synchronize the current IMU data with the delayed visual data (delay of about tens of milliseconds).

4.5.4. *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 and rather anatomically correlated. We thus formulate some anatomical constraints of the human hand as the measurement equations to be used in the EKF correction stage. 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. Total six of such positional constraints are applied (CMC/MCP joints for the thumb; MCP/PIP joints for the index and middle fingers). We also define the rotational constraint (e.g., PIP joint cannot rotate about the x-axis (i.e., no twisting)). Total seven of such rotational constraints are applied (no x-axis rotation of MCP joints for the three fingers; no x/z-axes rotations of the PIP joints for the index and middle fingers) following the adopted anatomical model (Fig. 4.1).

The observation matrix for both the positional and rotational constraints is derived as follows. First, the positional constraint is defined to force adjacent segments to be connected at the pivot joints (e.g., the intermediate and proximal phalanges are connected at the PIP joint), which means that the global position of the connecting joint is the same when estimated from the adjacent segments. The measurement equation of the positional constraint for the joint a, which connects the adjacent segments k and l, is given as

$$\begin{aligned} \tilde{z}_{J_a,p} =& \ z_{J_a,p} - \hat{z}_{J_a,p} = -\hat{z}_{J_a,p} \\ \hat{z}_{J_a,p} \ =& \ \hat{p}^{G}_{G,I_k} + \hat{R}^{G}_{G,I_k}\hat{R}^{I_k}_{I_k,B_k}d(\hat{\lambda}_{B_k})L^{B_k}_{B_k,J_a} - (\hat{p}^{G}_{G,I_l} + \hat{R}^{G}_{G,I_l}\hat{R}^{I_l}_{I_l,B_l}d(\hat{\lambda}_{B_l})L^{B_l}_{B_l,J_a}) \end{aligned} \tag{4.32}$$

where $z_{J_a,p}$ has always zero value since this measurement is a virtual measurement forcing the two segments connected on the pivot joint. Thus, $\tilde{z}_{J_a,p}$ is same with the estimated residual of the constraint $\hat{z}_{J_a,p}$ with the negative sign.

The observation matrix for this measurement $H_{J_a,p}$ is defined by the linearization of this measurement equation. Note that, different from the marker measurement equation, this measurement equation is related to the states of the both segments k and l at the same time, which results in the block matrices $H_{J_{a,k},p}$ for x_k and $H_{J_{a,l},p}$ for x_l. The block matrix $H_{J_{a,k},p}$ for the segment k is given by

$$H_{J_{a,k},p} = [H_{\tilde{p}^G_{G,I_k}} \quad 0_{3\times3} \quad H_{\tilde{\theta}^G_{G,I_k}} \quad 0_{3\times6} \quad H_{\tilde{\theta}^{I_k}_{I_k,B_k}} \quad H_{\tilde{\lambda}_{B_k}}] \tag{4.33}$$

$$\begin{aligned} H_{\tilde{p}^G_{G,I_k}} &= -I_{3\times3} \\ H_{\tilde{\theta}^G_{G,I_k}} &= \hat{R}^G_{G,I_k}[\hat{R}^{I_k}_{I_k,B_k} d(\hat{\lambda}_{B_k}) L^{B_k}_{I_k,M_j} \quad \times] \\ H_{\tilde{\theta}^{I_k}_{I_k,B_k}} &= \hat{R}^G_{G,I_k}\hat{R}^{I_k}_{I_k,B_k}[d(\hat{\lambda}_{B_k}) L^{B_k}_{I_k,M_j} \quad \times] \\ H_{\tilde{\lambda}_{B_k}} &= \hat{R}^G_{G,I_k}\hat{R}^{I_k}_{I_k,B_k} d(L^{B_k}_{I_k,M_j}) \end{aligned} \tag{4.34}$$

The block matrix $H_{J_{a,l},p}$ for the segment l has a much similar form with the $H_{J_{a,k},p}$, which also can be defined by partially differentiating the measurement equation.

Second, the other constraint is the rotational constraint (i.e., the PIP joint cannot be rotated about the x-axis (twisting)). whose measurement equation is given as

$$\begin{aligned} \hat{z}_{J_a,q} &= \beta \; q^{B_k}_{B_l} = 0 \\ q^{B_k}_{B_l} &= ((q^{I_k}_{B_k})^{-1} \otimes (q^G_{I_k})^{-1} \otimes (q^G_{I_l}) \otimes (q^{I_l}_{B_l})) \end{aligned} \tag{4.35}$$

where $\otimes$ represents quaternion multiplication, and β is a 1×4 row basis vector representing the nonrotatable axis (e.g., $\beta = [1,\ 0,\ 0,\ 0]$ when joint J_a has no DOFs in the x-axis).

The observation matrix for this measurement, $H_{J_a,q}$ is also defined by linearization for the both segments k and l, which results in the block matrices $H_{J_{a,k},q}$ and $H_{J_{a,l},q}$. The block matrix $H_{J_{a,k},q}$ for the segment k is given by

$$H_{J_{a,k},q} = [0_{3\times6} \;\; H_{\hat{\theta}^{G}_{G,I_k}} \;\; 0_{3\times6} \;\; H_{\hat{\theta}^{I_k}_{I_k,B_k}} \;\; 0_{3\times3}] \tag{4.36}$$

$$\begin{aligned} H_{\hat{\theta}^{G}_{G,I_k}} &= \beta \; [\hat{q}^{B_k}_{I_k}]_L \; [\hat{q}^{I_k}_{B_l}]_R \begin{bmatrix} -I_3/2 \\ 0 \end{bmatrix} \\ H_{\hat{\theta}^{I_k}_{I_k,B_k}} &= \beta \; [\hat{q}^{B_k}_{B_l}]_R \begin{bmatrix} -I_3/2 \\ 0 \end{bmatrix} \end{aligned} \tag{4.37}$$

Similarly, the block matrix $H_{J_{a,l},q}$ for the segment l is given by

$$H_{J_{a,l},q} = [0_{3\times6} \;\; H_{\hat{\theta}^{G}_{G,I_l}} \;\; 0_{3\times6} \;\; H_{\hat{\theta}^{I_l}_{I_l,B_l}} \;\; 0_{3\times3}] \tag{4.38}$$

$$\begin{aligned} H_{\hat{\theta}^{G}_{G,I_l}} &= \beta \; [\hat{q}^{B_k}_{I_l}]_L \; [\hat{q}^{I_l}_{B_l}]_R \begin{bmatrix} I_3/2 \\ 0 \end{bmatrix} \\ H_{\hat{\theta}^{I_l}_{I_l,B_l}} &= \beta \; [\hat{q}^{B_k}_{B_l}]_L \begin{bmatrix} I_3/2 \\ 0 \end{bmatrix} \end{aligned} \tag{4.39}$$

where the $[q]_L$ and $[q]_R$ are respectively the left and right quaternion product matrices.

4.6. Quantitative Evaluation for Free Hand Motion

4.6.1. *Experimental Setup*

We evaluate the tracking performance of our proposed VIST framework for the free hand motion, which is a standard scenario for evaluation adopted by many other results. For the quantitative evaluations, we employ Optitrack® MOCAP system and attach IR reflective markers on the four keypoints (Fig. 4.6-A). Subjects are instructed to sit on a chair surrounded by the MOCAP camera, where the stereo camera is positioned in front of the subject on a table facing downward at about 60 degrees. The camera is fixed during the experiment so that the tracking errors of the hand motions relative to the camera $\{C\}$ can be purely measured. This is because most existing vision-based studies also track their hand motions with respect to the camera (Yang & Ramanan 2012; Mueller *et al.* 2018; Moon *et al.* 2018; Iqbal *et al.* 2018).

We attach motion capture (MOCAP) markers to four keypoints of the glove (three at the fingertips and one on the hand dorsum). We choose the key points in this way, since: 1) fingertips typically exhibit larger tracking error than finger midpoints (e.g., MCP, DIP, or PIP (Mueller *et al.* 2018; Moon *et al.* 2018)), thus, smaller errors with our keypoints would imply better tracking even if the measured points are different; and 2) our MOCAP system cannot track robustly more than four markers, as they are anonymous and moving within a small size space (although our VIST system can handle a much larger number of markers thanks to its TC-fusion).

The participated fifteen subjects are all right-handed in the age range of 22 to 31 years, with no known perception disorders and have various hand shapes, as shown in Fig. 4.7. The experiments are conducted in accordance with the requirements of the Helsinki Declaration.

4.6.2. *Procedure*

As shown in Fig. 4.6-A, a subject is instructed to sit on a chair surrounded by the MOCAP camera, and to duplicate a hand configuration randomly chosen from a large hand image set (Mueller *et al.* 2018; Zhang

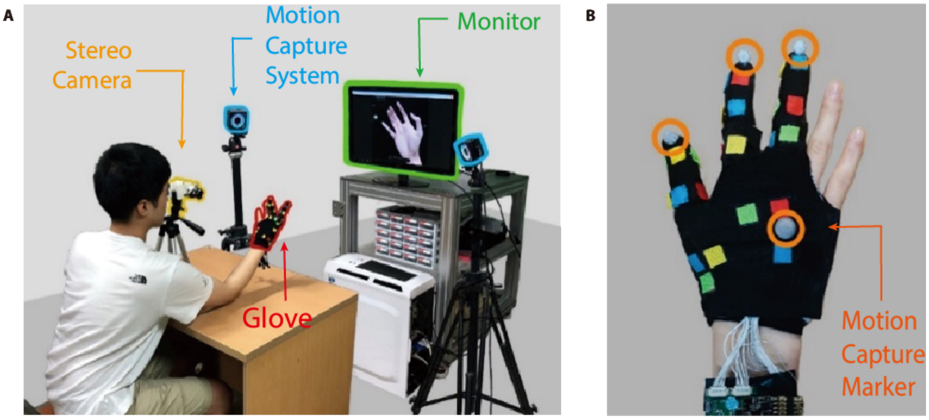


FIGURE 4.6. **(A)** The setup for the quantitative experiment. The subjects are instructed to follow randomly-displayed hand images on the monitor. The tracking error of the keypoints are measured from the surrounding MOCAP system. **(B)** The attachment configuration of MOCAP markers on the sensor glove.

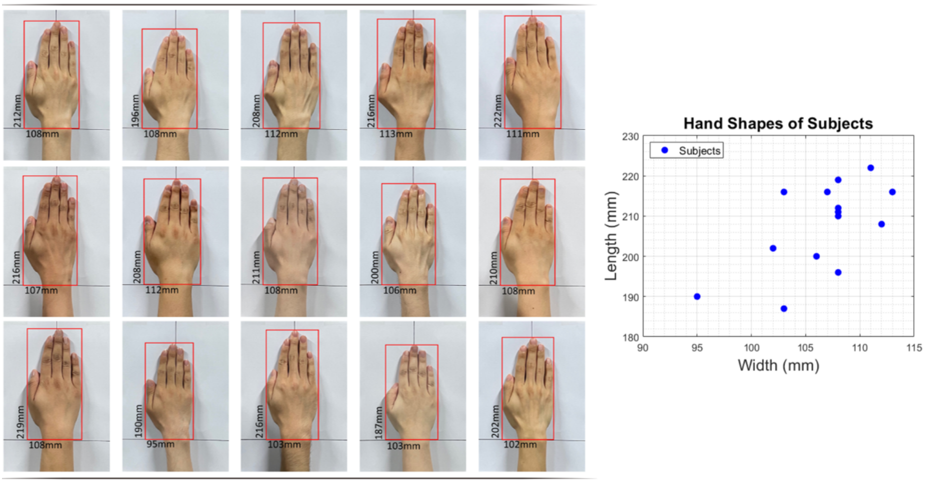


FIGURE 4.7. Snapshots of the hand shapes of all the subjects with their length from wrist to middle finger-tip and the width from thumb MCP to right side of the hand, which are plotted with respect to the length and width.

et al. 2017) displayed on the monitor every 3 seconds during 5 minutes. The displayed images were randomly selected from two large datasets of synthetic hands (Mueller *et al.* 2018) and real hands (Zhang *et al.* 2017) Since these images are adopted from other vision-based methods (Mueller *et al.* 2018; Iqbal *et al.* 2018) for evaluations. By deploying these datasets, we indirectly compare our results with other up-to-date systems. The time interval for the next random image are decided as 3 seconds through a pilot test, where if the time interval between random images is too short, subjects cannot follow the displayed images correctly, while too long time interval leads to an underestimation of the tracking error. We found that this time interval is an adequate time interval for subjects to follow the referenced images correctly and quickly at the same time.

4.6.3. *Experimental Result*

We analyze the tracking performance with two metrics: mean tracking error and percentage of correct keypoints (PCK), which is the percentage of the frames from all the frames, for which the maximum tracking error of all the keypoints is within a certain error threshold (Yang & Ramanan 2012). This PCK is a popular criterion adopted by many hand tracking systems (Armagan *et al.* 2020; Mueller *et al.* 2018; Moon *et al.* 2018; Zimmermann & Brox 2017).

As shown in table 4.1, the mean errors of all subjects for each joint are measured as 8.93 mm, 11.08 mm, 11.87 mm, and 10.89 mm, and the mean value of these errors (i.e., the tracking error of VIST) is obtained as 10.69 mm. We also compute percentage of correct keypoints (PCK) metric, which is the percentage of the frames (scenes) from all the frames, for which the maximum tracking error of all the keypoints is within a certain

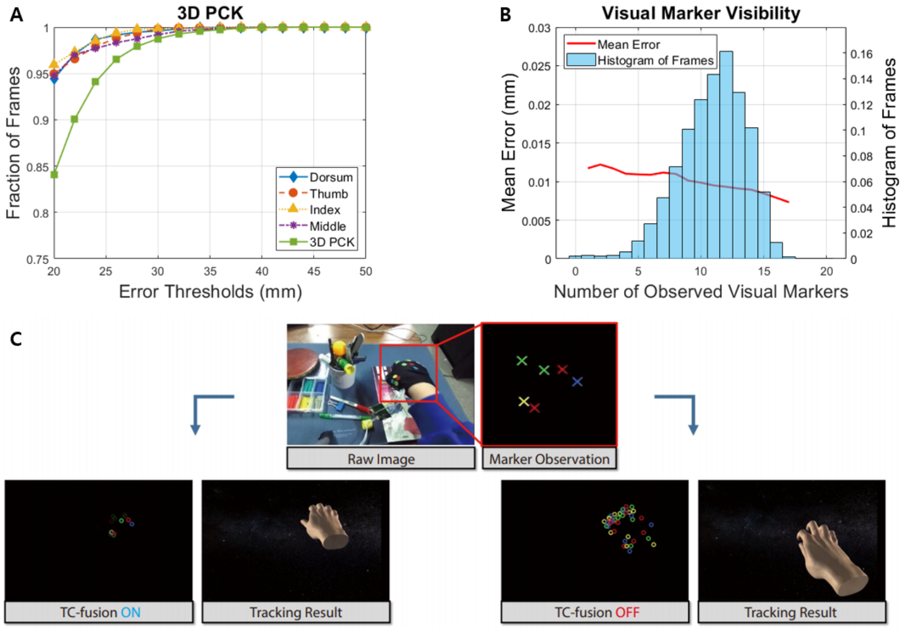


FIGURE 4.8. **(A)** The 3D PCK analysis of each keypoint. **(B)** The histogram of frames (blue) and the mean error (red) with respect to the number of observed markers. **(C)** Unstable tracking with TC-fusion (i.e., IMU-aided correspondence search) deactivated.

error threshold (Yang & Ramanan 2012). This PCK is a popular criterion adopted by many hand tracking systems (Armagan *et al.* 2020; Mueller *et al.* 2018; Moon *et al.* 2018; Zimmermann & Brox 2017). The PCK metric of our system is obtained as 84 % within 20 mm error, and about 99 % within 35 mm (Fig. 3B). The SDs of the tracking errors are only about 1.57 mm (between subjects) and 0.41 mm (for overall data) as shown in (Fig. 3C), despite of different hand shapes of all subjects (Fig. S6). Given that individually-different hand shapes are one of the most challenging issues in vision-based systems (i.e., larger error for hand shapes not in the training set (Armagan *et al.* 2020)), our small value of SD certifies that our VIST clearly overcome the generalization issues of existing vision-

Keypoint (Free Motion)	Mean Joint Error	SD (Subjects)	SD (Overall Data)	95 % CI for Mean Error
Dorsum	8.93	1.74	0.45	[8.05 , 9.81]
Thumb	11.08	1.69	0.44	[10.22 , 11.93]
Index	11.87	1.52	0.39	[11.10 , 12.64]
Middle	10.89	1.34	0.35	[10.21 , 11.57]
Mean	10.69	1.57	0.41	[9.90 , 11.49]

TABLE 4.1. The experimental results with statistical analysis of each keypoint from the quantitative evaluation for free hand motion. The mean joint error, standard deviations (SD) between subjects and overall data, and 95 % confidence interval (CI) for mean error.

based approaches. The absolute errors and percentage errors along the lengths of the hand lengths are also shown, which also verifies the robustness/accuracy of our VIST regardless of the hand shapes. The 95 % CI of the mean error is obtained from 9.90 mm to 11.49 mm and this narrow width of the CI certifies that the experimental results are statistically confident.

Moreover, in comparison to the latest hand-tracking challenge (Armagan *et al.* 2020), our VIST shows slightly improved metric than the winner of the challenge (13.66 mm (Zhang *et al.* 2020)) among RGB-D vision-based tracking systems, which have limitations of demanding computation, heavier hardware, higher power consumption, and limited-outdoor usage as compared to our framework. The mean error is computed as 10.14 mm, which is much smaller than state-of-the-art RGB-camera-only methods (about 50 mm (Mueller *et al.* 2018)). The PCK metric of the

VIST framework also can be compared with those of the state-of-the-art vision-based systems (Mueller *et al.* 2018; Zhang *et al.* 2020). The PCK of our system within 20 mm error is obtained as 80 %, which is higher than the compared algorithms, where the PCKs are only about 45 % or less at the 20 mm error threshold. These fairly improved PCK metric implies the enhanced user experience of the tracking system in terms of human perception as presented in the previous chapter 2. The more discussion in terms of human perception of tracking accuracy of our VIST will be addressed in the following discussion section.

We present the mean error and histogram for frames with respect to the number of observed markers (Fig. 4.6B). The mean error does not substantially increase as the number of visible markers decreases (except the rare case where only one marker is observed), as opposed to general vision-based systems wherein the tracking errors sharply deteriorate with occlusions (Armagan *et al.* 2020; Moon *et al.* 2018; Mueller *et al.* 2018). This robustness against occlusions is because our VIST framework can still estimate the occluded segments by opportunistically exploiting the IMU information. Note also that the histogram of frames with respect to the number of observed markers clearly follows a normal distribution, indicating that the displayed hand images in the experiment are not biased. This demonstrates that the robustness of our system is applicable to not only certain/specific but also general/diverse postures of the hand.

We also perform experiments with the TC-fusion off and the hand tracking simple becomes unstable/chattering due to the issue of correspondence search as stated in chapter 4.3 (Fig. 4.6C); and evaluate the stereo camera and the IMU-only EKF individually and they show much poorer performance than our VIST (i.e., camera: mean error 12.7 mm

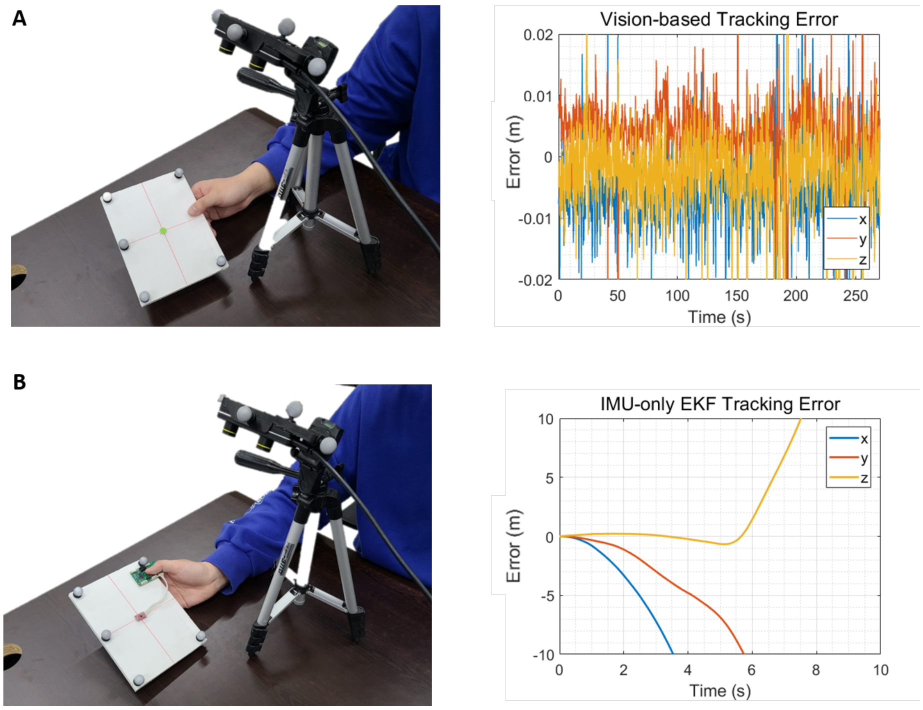


FIGURE 4.9. Experimental setup (left) for measuring the positional accuracy and tracking error (right) of each sensing modality. **(A)** The mean error from the visual marker/sensor is computed as 12.70 mm, which is larger than 10.14 mm of the proposed VIST algorithm, **(B)** The mean error estimated from IMU-only EKF drastically diverges in seconds owing to uncalibrated parameters of the IMU (drift, bias, noise, or disturbance).

with 30Hz rate; IMU-only EKF quickly diverging) even for merely a single point tracking (Fig. 4.9). These all clearly manifest the importance/cruciality of the TC-fusion for our proposed VIST framework.

During the experiment, runtime performance of the proposed algorithm is also measured as shown in table 4.2. The frequency (the number of iterations per second) and the execution time per iteration of each subprocess of the proposed VIST algorithm in real-time operation. Mainly

Subprocess	Frequency [Hz]	Execution Time [μs]
IMU Data Acquisition	91	9374
Prediction	"	1210
Images Acquisition	26	8825
Marker Detection	"	2931
Stereo Matching	"	85
Correspondence Search	"	93
Correction	"	2021

TABLE 4.2. The Runtime performance of the proposed algorithm.

two threads (an IMU thread (up) and a vision thread (down)) are running, where the raw data acquisition processes from each sensor (including waiting time for arrival of next-step sensor data) account for most of the execution times, which shows the computational affordability of our algorithm. All the process only takes up less than 25% CPU usage in Intel i7-7700HQ/2.80 GHz laptop without requiring high-end GPU.

4.7. Quantitative and Comparative Evaluation for Challenging Hand Motion

4.7.1. *Experimental Setup*

Using the same setup in Sec. 4.6, we conduct quantitative experiments for the two scenarios, object-interaction and wearing haptic devices, which raise challenging issues of severe occlusion, magnetic interference, or mechanical contact. Our experiments aim to verify the performance/robustness of the tracking algorithms for general/various hand motions, thus, we generate large instruction sets of hand images

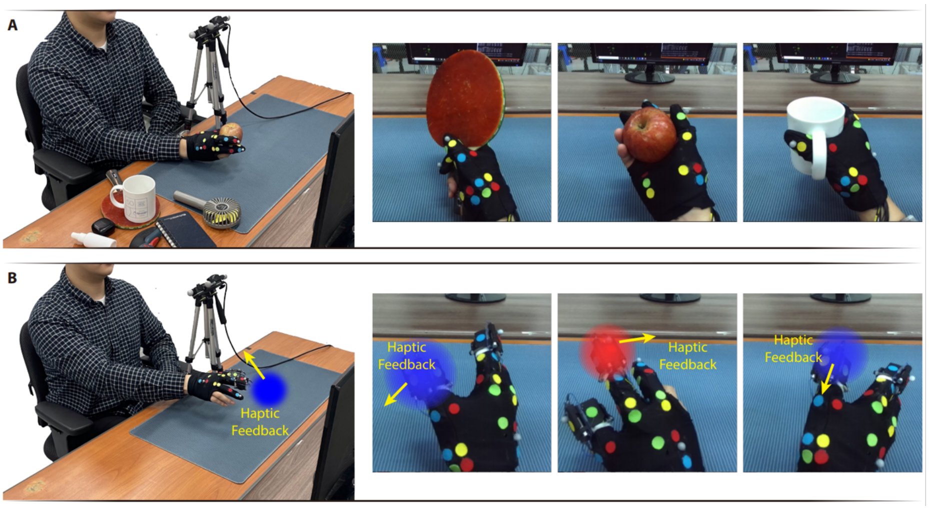


FIGURE 4.10. Quantitative evaluation setup for **(A)** object-interaction, which display random object-interaction images from instruction set and **(B)** wearing haptic devices, which display virtual circle instruction where user can receive haptic feedback via CHDs.

(for displaying random hand images) and assign all different instruction sequences for each subject. The experiments are performed using the setup with random instruction (Fig. 4.10, the detailed description of which would be presented in following section.

4.7.2. *Procedure*

In case of the experiment for object-interaction, we construct a large instruction set of diverse objects (i.e., an apple, mug, cosmetic, racket, book, earphone case, portable fan, and hand-drill). Those objects are selected as experimental objects, since we frequently interact in daily life. The dataset of each object consists of hundred of hand images (i.e., eight hundred images for the instruction set) including possible configurations of hands when interacting with the object. The experiment for a sub-

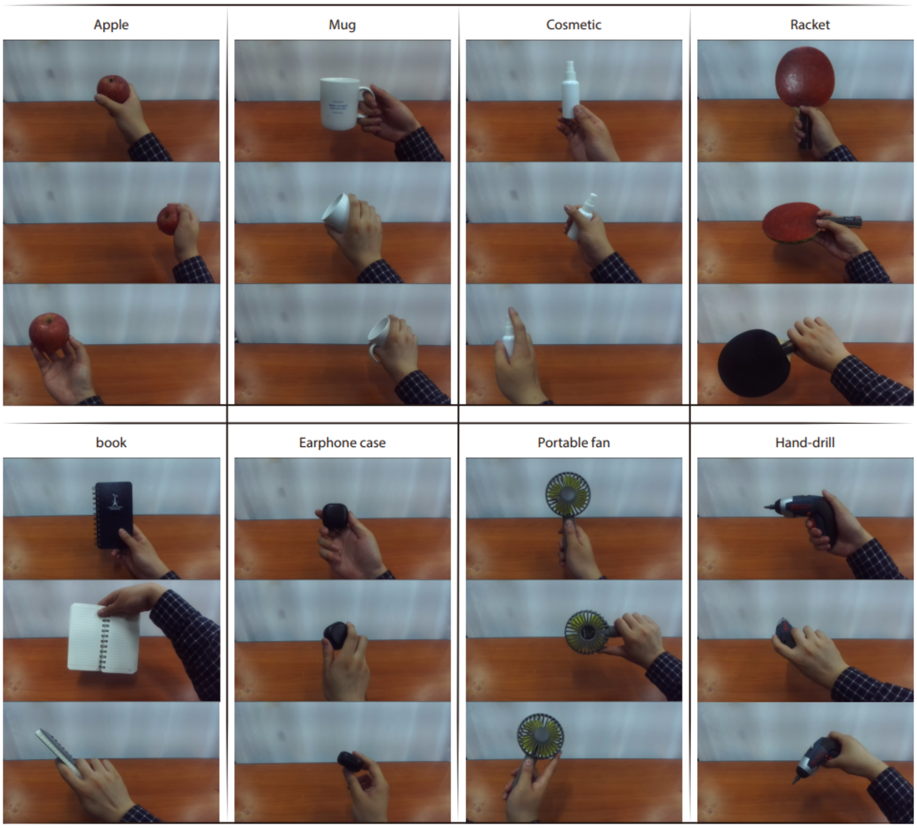


FIGURE 4.11. Snapshots of an instruction set for object-interaction experiment. The set includes images of various hand poses interacting with diverse daily objects including mangetic objects (earphone case, portable fan, and hand-drill). The instruction set of each object consists of hundred of hand images.

ject consists three sets, where the target objects are selected randomly among instructional objects (Fig. 4.10A). In each set, the subject conducts ten trials of hand motions, where, same with Sec. 4.6, the subject is instructed to duplicate a displayed image on the monitor, which is randomly sampled every three seconds from the dataset of the target object. The time interval between trials is three seconds, and, including transitional motions to the target object of the next set, the duration of each set is forty seconds, thus, the total duration of each experiment is two minutes. To obtain robust/generalizable experiments results, total fifteen subjects participate in experiments with various hand shapes (Fig. 4.7) same with the previous experiment. All the number of data is amount to 45K (about 3K data for each subject) for each experiment, which certify statistically significance of our experiments

In case of the experiment for wearing haptic devices, we construct an instructional program where the range of haptic feedback is randomly displayed (Fig. 4.10B). In each trial, the range of haptic feedback is displayed upon the raw left images real-time, which take forms of a translucent circle and the location/radius of the circle is randomly determined. When displaying the haptic range, the subjects are asked to their finger-tips to the center of range. Through the CHDs, the haptic feedback is delivered on the finger-tip, where the magnitude of the feedback increases as the finger-tips are closer to the center of the circle. The time interval between trials is also three seconds and each subject conducts forty trials of hand motions, thus, the total duration of each experiment is also two minutes, which amount to about 30K data and fifteen participants participate in the two experiments same with Sec. 4.6.

4.7.3. *Experimental Result*

The experimental results with statistical analysis are shown in table 4.3 and 4.4. The mean errors (SDs) of all joint errors are measured as 12.68 mm (1.66 mm) for the object-interaction experiment (table 4.3) and 10.89 mm (1.53 mm) for the haptic device experiment (table 4.4), and the PCKs of all joints are presented comparing with the previous results of free hand motion (Fig. 4.12). The 95 % CIs of the mean errors

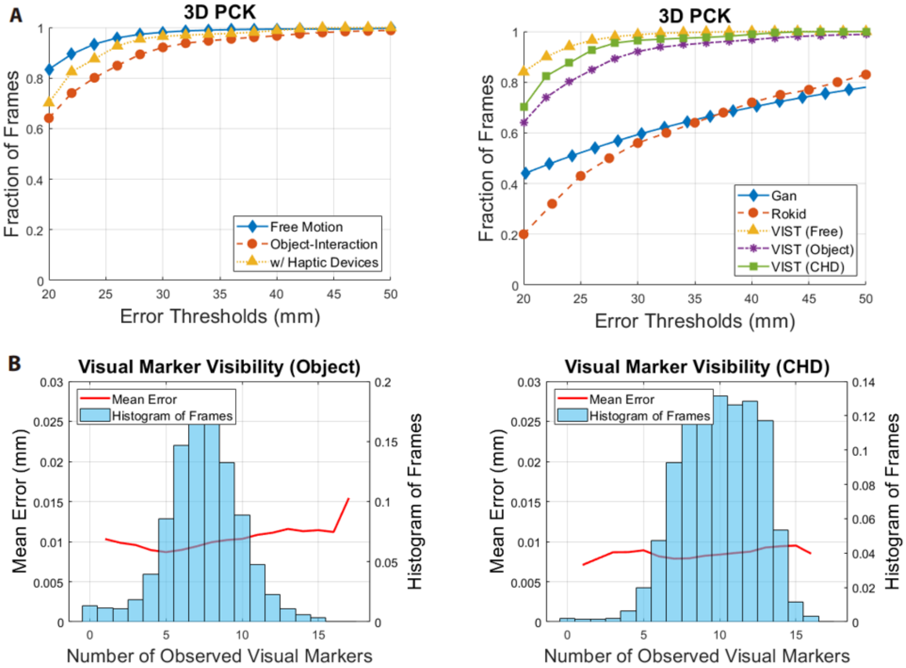


FIGURE 4.12. **(A)** 3D PCK analysis of the experiments (object-interaction and wearing haptic devices) in comparison with the previous free motion result (left) and other tracking methods for free motion (Iqbal *et al.* 2018; Zhang *et al.* 2020), which is the winners of the tracking challenge (Armagan *et al.* 2020). **(B)** The histogram of frames (blue) and the mean error (red) with respect to the number of observed markers.

are computed in narrow ranges, by [11.36 mm, 13.76 mm] and [9.82 mm, 12.51 mm] respectively, which verify that the obtained mean errors are statistically significant for general users.

For the experiment of object-interaction, compared to the winner of the RGB vision-based hand tracking with objects (Iqbal *et al.* 2018) in the challenge (Armagan *et al.* 2020), our VIST system shows the much improved PCK metric and mean errors. Given the difficulty of even detection of the bounding box of the human hand, in case of large occlusion

Keypoint (Object-Interaction)	Mean Joint Error	SD (Subjects)	SD (Overall Data)	95 % CI for Mean Error
Dorsum	9.85	1.51	0.39	[9.08 , 10.61]
Thumb	12.29	1.88	0.48	[11.34 , 13.24]
Index	14.04	1.54	0.40	[13.26 , 14.82]
Middle	14.52	1.71	0.44	[13.66 , 15.39]
Mean	12.68	1.66	0.43	[11.84 , 13.52]

Table 4.3. The experimental results with statistical analysis of each keypoint from the quantitative evaluation for object-interaction.

Keypoint (w/ Haptic Device)	Mean Joint Error	SD (Subjects)	SD (Overall Data)	95 % CI for Mean Error
Dorsum	9.14	1.64	0.42	[8.31 , 9.97]
Thumb	10.82	1.52	0.39	[10.05 , 11.59]
Index	12.22	1.41	0.36	[11.50 , 12.93]
Middle	11.38	1.56	0.40	[10.59 , 12.17]
Mean	10.89	1.53	0.40	[10.11 , 11.66]

Table 4.4. The experimental results with statistical analysis of each keypoint from the quantitative evaluation for wearing haptic devices.

(thus, the challenge also provides the ground-truth position of the hand), these contrasting results can verify that superior performance/robustness of our VIST with objects. In case of wearing haptic devices also cause occlusion by wearing haptic devices, which leads to the visual distortion of human hands from the dataset for the vision-based systems.

Note that the CHDs and the three instructional objects (i.e., hand-drill, earphone case, and portable fan) comprise of ferromagnetic materials (e.g., steel, magnets) or generate operating current, which cause electro-magnetic interfere to magnetometers. Thus, tracking hands wearing the devices or interacting with those objects (e.g., electronic devices) are challenging for IMU/compass wearable systems, for the same hand motions from the VIST experiment, while our VIST exhibits the robust performance for those scenarios as mentioned earlier. In summary, through these experiments, we can again verify the superior performance and robustness of our proposed VIST system and its promise for many challenging real-world applications.

4.7.4. *Performance Comparison with Existing Methods for Challenging Hand Motion*

In order to compare the different algorithms fairly, we think, by doing so the superior performance of our VIST can clear be shown, the same dataset of hand motions should be applied to the different algorithms. Unfortunately, feeding the exactly same raw data for the different methods (i.e., vision-based tracking or IMU/compass wearable tracking) with our VIST data is impossible, due to the sensor configuration is completely different. To be more specific, the training sets of vision-based systems mostly are generated from bare hand images, yet, the hand images from our experiment wearing the sensor glove, thus, cannot be directly applied to vision-based methods. In addition, a strength of the proposed VIST is free from the problematic magnetometers, thus, the wearable sensor glove is fabricated only with the accelerometers and gyroscopes, which also allows for improvement of sampling rate of our sensor glove.

Especially, compared to the previous experiments of free hand motion, which extract instructional images from one dataset of the compared method (Zhang *et al.* 2019*a*), thus the result can be generalizable only for the dimension of hand configuration. On the contrary to this, for the challenging scenarios, the dimension of the hand motions are increased (e.g., for various types of objects) and the similar dataset are absent for the existing methods (e.g., wearing CHD). Thus, although exact same hand motion is not tested, we alternatively conduct comparative study for the same instructional motions to show that our VIST outperforms other methods for the challenging scenarios. We gather datasets of hand motions as similar as possible for same subjects, and compare the different algorithms by feeding the similar datasets.

A public library of vision-based tracking (Zhang *et al.* 2019*a*) is tested for the object-interaction (book, racket, and mug) and wearing CHD. The tracking error of the method is also measured with the MOCAP system, by attaching four MOCAP markers on the same position with the previous experiment, and calibrating it from the vision-based tracking results. The tracking results are shown in Fig. 4.13. To compare with our VIST results, the hand motions from the same instruction in the experiment (Sec. 2.4.5) is tested. When the hands are interacting with the objects or wearing the haptic devices, the tracking results (green line) becomes much unstable (distorted from real hand). The measured

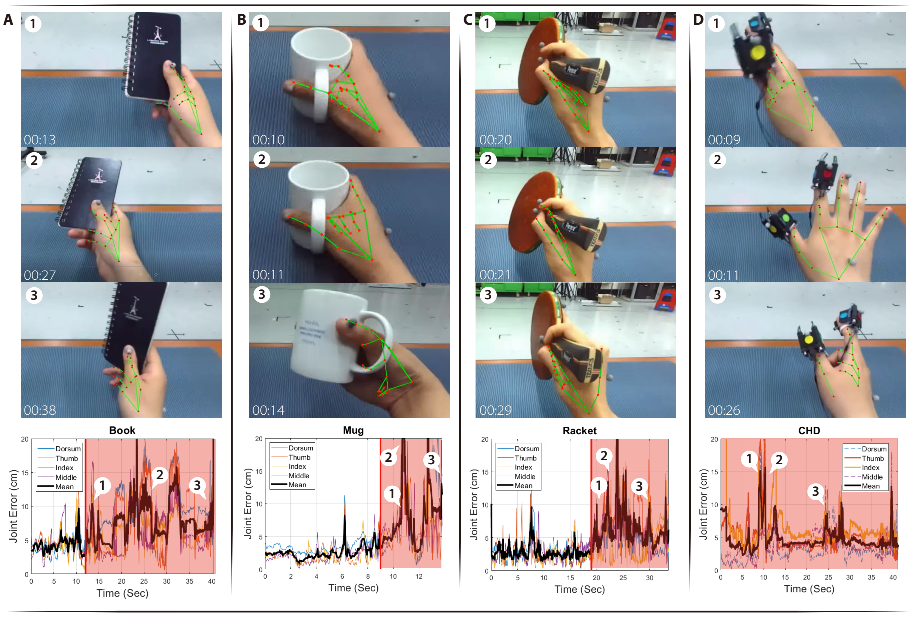


FIGURE 4.13. Snapshots (top) and tracking error obtained by MOCAP (bottom) of a vision-based method [9] in cases of object-interaction (**(A)** hand-drill, **(B)** mug, **(C)** racket) or **(D)** wearing haptic devices (errors of thumb/index fingers (yellow, red) are explicitly increased due to the presence of CHDs).

tracking error (black) is reasonable at some extent in free motion, but when self-occlusion occurs, the tracking error largely increased from the stable result. Moreover, when the hands are interacting with the objects or wearing the haptic devices (red area), the tracking error sharply increased (even up to tens of *cm* more than their hand size), which clearly shows the limitations of the vision-based method to occlusion.

We alternatively conduct comparative study for the same instructional motions to show that our VIST outperforms other methods for the challenging scenarios. For the case of magnetic-interference, we also conduct the running test for the same instructional most of VIST. Then we save all the calibrated sensor data of compass, adopting the same sen-

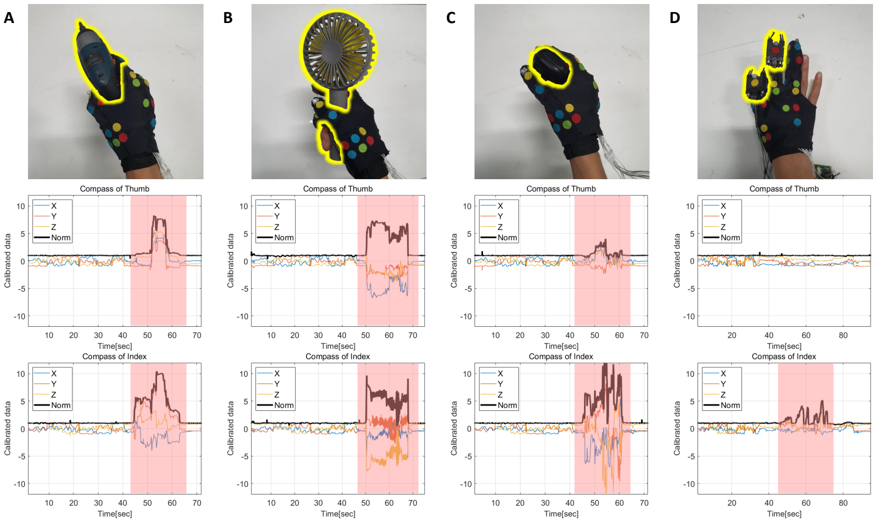


FIGURE 4.14. Snapshots (top) and calibrated sensor data (middle, bottom) of the compasses (on the metacarpal phalanx of the thumb finger (middle) and the proximal phalanx of the index finger (bottom)) when cases of object-interaction (**(A)** hand-drill, **(B)** portable fan, **(C)** earphone case) or **(D)** wearing haptic devices.

sor configuration with the FTM in the previous chapter 3. As explained earlier, the compass should be pre-calibrated to provide global yaw information by measuring earth magnetic field, thus, after the calibration the magnitude of compass reading should be one (unit vector for directional information). However, as shown in Fig. 4.14, the magnitude of the calibrated reading data (black line) sharply increased due to the strong magnetic field from objects (hand-drill, portable fan, earphone case) or CHDs. This means the compass data would hamper the rotational estimation when the hands wear the haptic devices or interacting with magnetic objects (e.g., electronics, still-wall, magnets), which shows the unstable performance of IMU/compass wearable systems.

4.8. Qualitative Evaluation for Real-World Scenarios

We perform qualitative evaluations of our VIST framework for real-world challenging scenarios, that defy existing hand tracking methodologies. These scenarios include the above representative challenging hand motions (i.e., object-interaction and wearing CHDs in Sec. 4.7), and other real-world scenarios, which can cause problems for existing methods (i.e., visually complex background (vision-based systems) or outdoor environment (both vision-based and wearable-sensor based systems). For this, we visualize the estimated hand pose/configuration in VR deploying the cross-platform game engine Unity® in the following figures.

4.8.1. *Visually Complex Background*

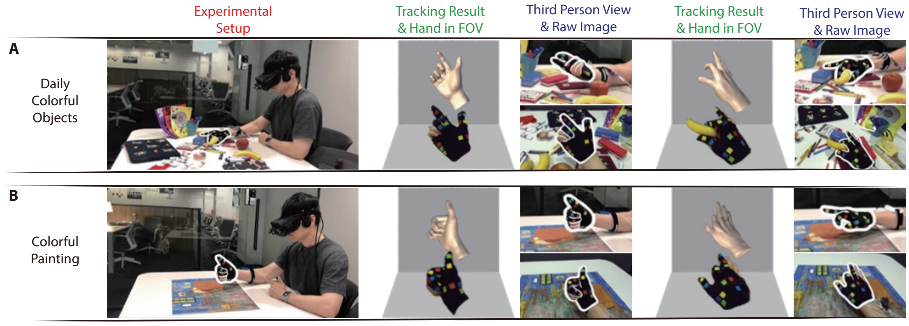

Experimental Setup
Tracking Result & Hand in FOV
Third Person View & Raw Image
Tracking Result & Hand in FOV
Third Person View & Raw Image
A
Daily Colorful Objects
B
Colorful Painting

FIGURE 4.15. Qualitative evaluations for visually complex background: **(A)** Colorful objects in the background or manipulated by hands and **(B)** Colorful painting background.

It is challenging for vision-based systems to track the human hand with objects or in the background with similar appearances/colors (Mueller *et al.* 2018; Zimmermann & Brox 2017; Sridhar *et al.* 2015). To evaluate robustness against such situations, we design qualitative experiments with colorful objects (magazines, fruits, stationery) and a painting ("Bedroom in Arles") in the background with visually similar colors/patterns to the glove markers. In addition to free hand motions, we also perform experiments of interaction with daily objects (banana and scissors).

As shown in Fig. 4.15, despite the visually adversarial objects/backgrounds, our VIST system robustly tracks the hand motion. This is because our VIST algorithm accurately detects visual markers from the background using the HSV and shape requirements (size, convexity, and circularity) simultaneously. Moreover, using the IMU-aided correspondence search, it can robustly eliminate outliers (i.e., color blobs in backgrounds with similar colors/shapes as the true markers) and match marker observations $\mathcal{O}$ with the IMU-predicted anonymous markers on the glove $\mathcal{M}$, thereby, ensuring this stable performance with visually complex objects/backgrounds.

4.8.2. *Object Interaction*

As noted earlier, occlusions are the fundamental limitations of the vision-based systems. Self-occlusion has sometimes been addressed, yet, still remains unsolved even in the state-of-the-art results (Moon *et al.* 2018; Mueller *et al.* 2017; Sridhar *et al.* 2016). When the hand is partially outside the FOV, the vision-based tracking cannot articulate the invisible segments or even recognize the human hand in a scene (Mueller *et al.* 2017, 2018). Therefore, we design a hand tracking experiment comprising a set of substantially self-occluded hand postures (e.g., fingers invisible behind palm or middle finger abduction behind index finger, which almost all vision-based tracking systems cannot handle). We also instruct the subjects to make various motions of the fingers substantially outside the FOV.

Our VIST system can track self-occluded hand poses accurately even

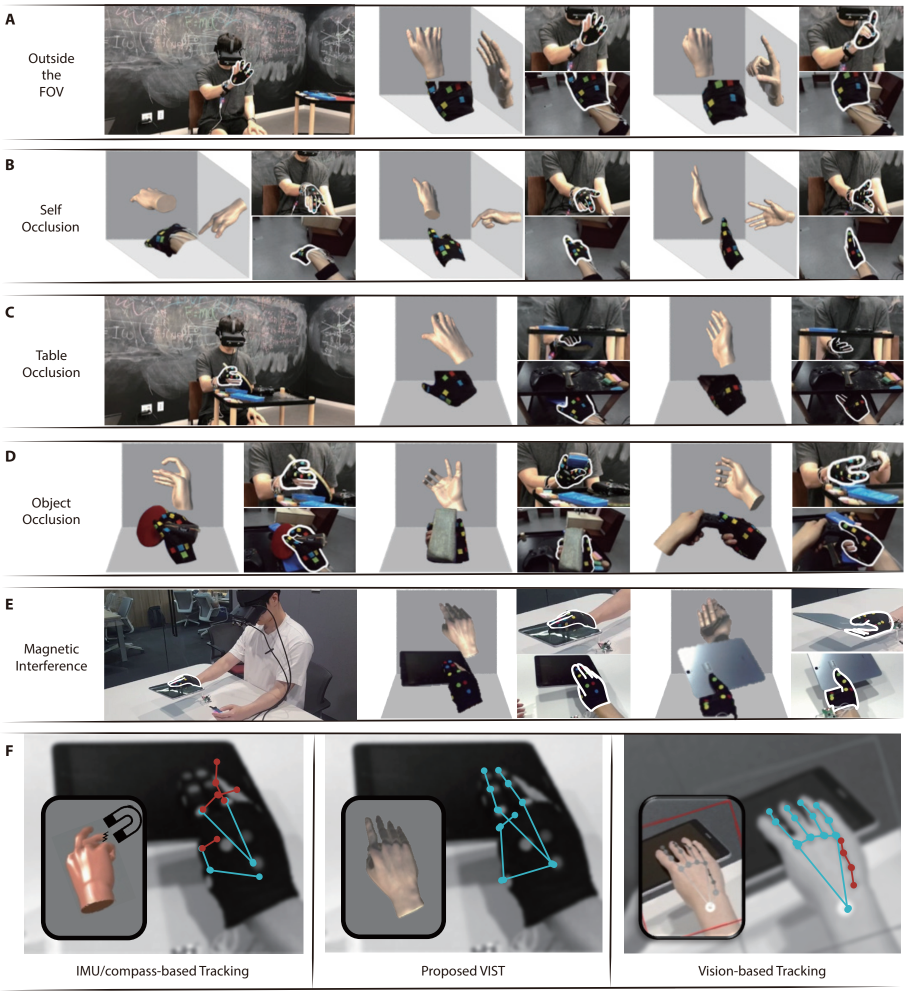


FIGURE 4.16. Qualitative evaluations for various occlusion types: **(A)** Outside the FOV, **(B)** Self-occlusion, **(C)** Severe occlusion from surroundings, **(D)** Interaction with various objects, **(E,F)** Robustness of VIST framework against magnetic-interference/contact/occlusion from the tablet and comparison with other tracking methods (Zhang *et al.* 2019*a*; Lee *et al.* 2019).

when half the hand is unobserved in the scenes (Fig. 4.16-B). This result clearly verifies the robustness of our system to self-occlusion and is consistent with the low error in the quantitative results regardless of the number of occluded markers. Moreover, our VIST system can track the poses of invisible segments outside the FOV (Fig. 4.16-A) as the IMUs with precisely real-time/auto-calibrated hand/sensor-related parameters can provide the required pose information. This property noticeably increases the usability of our system by allowing users to move their hands freely regardless of the FOV of the camera (e.g., for the camera used in our system, the operable area increases by about 50 % at a distance of 25 cm from the camera).

When human hands interact with objects, occlusion can be particularly challenging for the vision-based tracking. This is because it is infeasible to include all the daily objects with accurate annotations in the training set, or the tracking error typically worsens about 2–4 times for those untrained objects (Armagan *et al.* 2020). Even for objects in the training set, many systems fail to track hand motions with large objects (Mueller *et al.* 2017; Sridhar *et al.* 2016; Zimmermann & Brox 2017; Panteleris & Argyros 2017). Such object interaction is also problematic for soft wearable hand tracking, as the object may deform some soft sensors, which can then distort their signals related purely to the hand motion, thereby, causing bias or even instability in the estimated hand motion. On the other hand, the IMU/compass-based wearable tracking systems can be compromised in the case of interaction with objects containing ferromagnetic materials (e.g., metallic products, electronic components with magnets) or with electronic devices having substantial internal currents (e.g., powered tools, workstations), as the compasses of the systems are to be breached in this case.

In contrast, as shown in Fig. 4.16-C,D, our VIST framework retains accurate hand tracking even when the hand is occluded by, or interacting with, various objects (e.g., when behind the ping-pong racket or pressing the gampad buttons). Our VIST system can also track the hand under the table, thereby, verifying its robustness to occlusions from the surroundings. Additionally, we also conduct experiments to manipulate a tablet PC, which possesses embedded magnets and ferromagnetic materials (against IMU/compass tracking), form-factor prone to cause occlusions (against vision-based tracking) and contact on many parts of the hand (against soft-sensor tracking).

As shown in Fig. 4.16-E, our VIST system can still retain its tracking even with the tablet, whereas other existing systems fail as shown in Fig. 4.16-F.

4.8.3. *Wearing Fingertip Cutaneous Haptic Devices*

The vision-based hand tracking systems, which utilize datasets based on bare hands for the training, generally cannot track the hands when the user wears gloves or devices/attachments on the hand, as the appearance of the bare hand is then changed/distorted. The soft wearable tracking systems are also vulnerable to those extra devices/attachments as the soft sensor signals are distorted by their contacts. Large deformations due to the attachments can even cause permanent offset errors in the soft sensors. On the other hand, the IMU/compass wearable tracking systems can be severely interfered when the devices include magnets or actuators (Lee *et al.* 2019; Baldi *et al.* 2017).

To verify again the robustness of our VIST framework against such

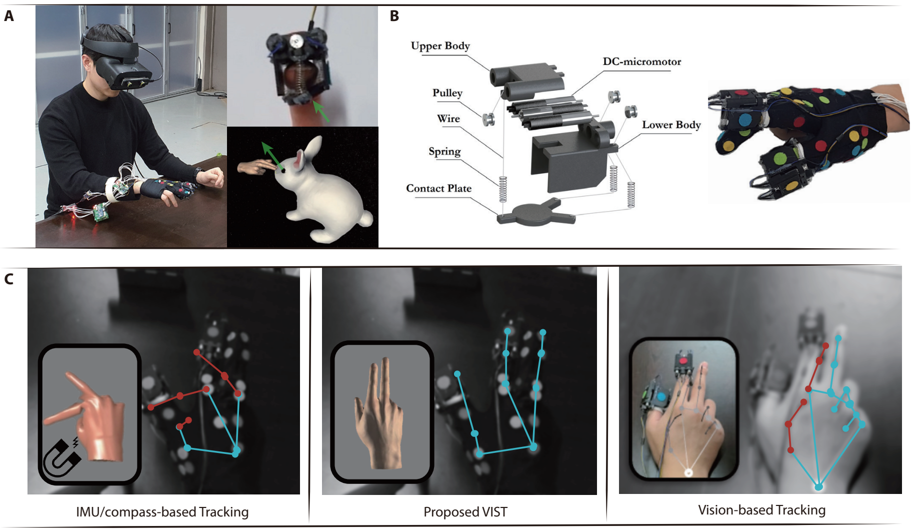


FIGURE 4.17. Qualitative experiments with CHDs in VR: **(A)** Experimental setup with virtual rabbit, **(B)** Construction of 3-DOF CHDs and wearing configuration, **(C)** Robustness of VIST framework against visual distortion, mechanical contacts and electromagnetic interference from CHDs and comparison with other tracking methods (Zhang *et al.* 2019*a*; Lee *et al.* 2019).

extra modules/gloves, we perform experiments with the CHDs same with the previous experiment, (Fig. 4.17B), which are developed to impose 3-DOF (i.e., sheer/normal) force feedback on the fingertip using three motors and a control plate for immersive VR experience. For the experiment, we also build a VR environment for haptic exploration task, where the human users can receive 3-DOF haptic feedback when touching a virtual rabbit (Fig. 4.17A). The users wear two CHDs on the thumb/index fingertips and the 3-DOF contact force is delivered on each of the fingertips. As shown in Fig. 4.17C and explained in the previous section 4.7, existing systems become unstable with the CHDs due to the visual distortion/occlusion of the hand or due to the magnetic interference from the magnets and current in the motors. In contrast, our proposed VIST

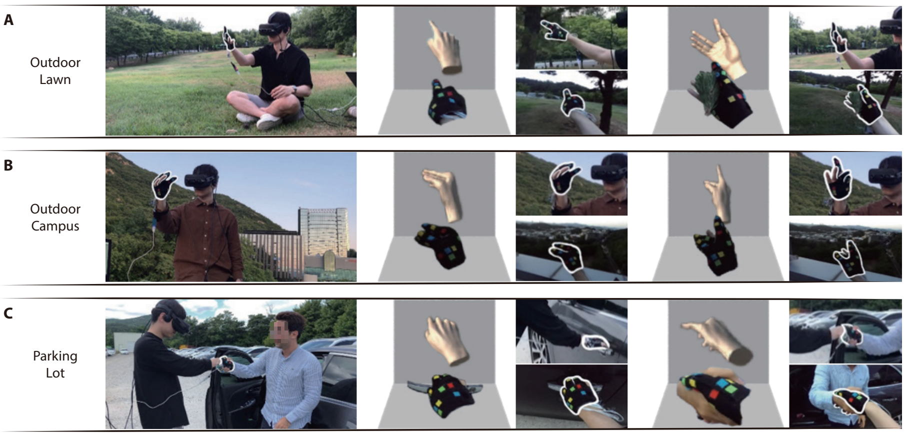


FIGURE 4.18. Qualitative experiments in outdoor environments: **(A)** Lawn, **(B)** Campus and **(C)** Parking lot.

system maintains the accurate hand tracking during the VR experiments.

4.8.4. *Outdoor Environment*

The IMU/compass or soft wearable hand tracking systems (Lee *et al.* 2019; Baldi *et al.* 2017; Glauser *et al.* 2019) typically require extra wrist position sensing, which is often based on IR-based tracker, that however is known to be affected by the ambient IR components in the sunlight. The RGB-D vision-based hand tracking systems are typically not suitable for outdoor environments either, since their structured IR rays can interfere with the sunlight as well (Meilland *et al.* 2015; Abbas & Muhammad 2012). For the outdoor operations, the RGB vision-based tracking systems also show lower accuracy, since their training sets are typically acquired indoors (Hampali *et al.* 2020; Zimmermann & Brox 2017; Zhang *et al.* 2017).

We thus design experiments in some challenging outdoor environments (i.e., lawn, campus, parking lot). In the outdoor lawn scenario, external power sources or powerful computing units (e.g., high-end GPUs (graphics processing units)), which are necessary for the external wrist sensors (for wearable hand tracking) or machine-learning techniques (for vision-based hand tracking), are unavailable. In the campus scenario, experiments are performed at sundown with drastically different light conditions from indoors, particularly detrimental for the RGB vision-based tracking. In the parking lot experiments, we include some everyday activities (opening car door, shaking hands), which are, yet, challenging for existing hand tracking methodologies owing to the ferromagnetic materials in vehicles (IMU/compass wearable tracking), mechanical contacts with people or objects (soft wearable tracking), or occlusions with different light condition (vision-based tracking). Our VIST system, in contrast, can robustly track the hand motion in all these experiments - see Fig. 4.18.

4.9. Discussion

Through the experiments, the superior performance of the proposed VIST frame work is quantitatively evaluated and verified. For the evaluation of free hand motion, the mean joint error is measured as 10.69 mm. Compared to the accuracy of other vision-based methods (e.g., 13.66 mm (Zhang *et al.* 2020)) of the winner of RGB-D vision-based tracking challenge (Armagan *et al.* 2020) or about 50 mm (Mueller *et al.* 2018) of recent RGB-only method), our obtained tracking accuracy shows the enhanced accuracy of VIST. The 3D PCK metic of all joint is measured as 84 % within 20 mm error, and about 99 % within 35 mm, which is fairly improved than those of existing systems The PCK metric of the VIST framework also can be compared with those of the state-of-the-art vision-based systems (Mueller *et al.* 2018; Zhang *et al.* 2020), which is higher than the compared algorithms, where the PCKs are only about 45 % or less at the 20 mm error threshold. Considering the PCK is the metric effectively how human user really perceive a tracking system, these results presents the superiority of our VIST in terms of human perception.

The efficacy of our VIST framework in terms of human perception can be more deduced reflecting the results of the previous chapter 2. In the chapter, when tracking the index fingertip in VR, human users cannot perceive tracking errors of the fingertip under 5 cm. Moreover, according to (Tan *et al.* 2007), they cannot discriminate angular difference of the index finger joints less than 1.7° based on the proprioception. Since the PCK of the VIST system within 5 cm error is 99 % and the rotation tracking error of the VIST system would be less than 1.57°, which is the rotational accuracy of purely IMU/compass-based tracking in (Lee *et al.* 2019) with no visual-inertial EKC/TC-fusion, our VIST framework would likely allow users to perceive their VR hands accurately following their real hands.

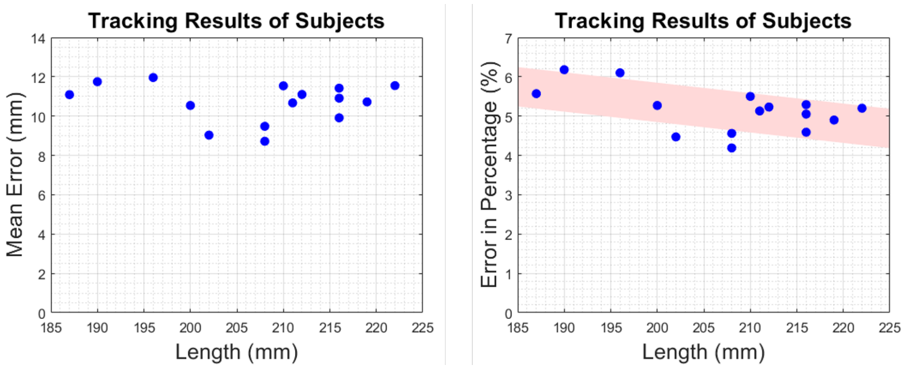


FIGURE 4.19. Tracking errors of each subject along with the their hand lengths. The absolute mean joint error (left), which is constant and the percentage error (i.e., mean joint error / hand length), which is downward slopping (linear regression result (red).

The SDs of the mean joint errors are computed as 1.57 mm (between subjects) and 0.41 mm (for overall data), 95 % CI are measured as 9.90 mm to 11.49 mm. This small value of SD and width of CI proves the statistical significance of our experimental results (i.e., generality to every user). Note that the tracking error is consistent with all different hands (Fig. 4.7) of the subjects although the hand size/shape is considerable factors of errors for existing systems (generalization issue for vision-based systems or accumulative error of forward kinematics for IMU/compass or soft wearable systems). On the contrary, as shown in the Fig. 4.19 (left), out VIST shows the consistency of absolute tracking error along with the hand size. This tendency can be more clear when the percentage error with the hand length is plotted (Fig. 4.19 (right)) where the linear regression line of the experimental data decreases.

In addition to free hand motion, we again verify the performance of VIST for challenging scenarios in quantitative manner (Sec. 4.7). These object-interaction case or wearing CHDs are still challenging difficult due to object-induced occlusion or magnetic interference. However, the mean error of VIST is still low value 12.68 mm for the object-interaction experiment and 10.89 mm for the haptic device experiment (table 4.3 and4.4). The results are also statistically meaningful given the small value of SDs (1.66 mm and 1.53 mm) and width of CIs for mean value ([11.36 mm, 13.76 mm] and [9.82 mm, 12.51 mm]), respectively. The gap of the tracking accuracy would become more distinct from the existing systems, which is natural since our VIST overcome the limitations of the systems. To present the superior performance of our system, we conduct a comparative studies of up-to-date vision-based method (Zhang *et al.* 2019*a*) and IMU/compass wearable system (Lee *et al.* 2019). For the hand motions of subjects/objects/instructions same with the VIST experiments, the systems shows sharply increased tracking error or unstable compass reading even after calibration, which exhibits the failure cases of existing systems compared to our VIST.

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, slow-update (Armagan *et al.* 2020; Mueller *et al.* 2017; Sridhar *et al.* 2016)), as 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/auto-calibrated 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, as the camera and IMUs are immune to them. Moreover, the integrated auto-calibration 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 taking several indicated poses, which is inevitably with human errors (Yuan *et al.* 2013; Luinge *et al.* 2007; Seel *et al.* 2014).

Chapter 5

Conclusion

As stated in the introduction, natural user interfaces based on hand motion tracking are promising for many current-burgeoning fields of human machine interaction such as robotics, VR/AR, rehabilitation, as we human mainly utilize hand to interact with objects and environments in daily life. Thus, in order to figure out the design specifications for hand tracking system, we study about the human perception of visual-proprioceptive conflicts (chapter 2), and suggest the design criteria (i.e., allowable error) for hand tracking system. Based on the findings, then, we develop the wearable finger tracking module (chapter 3) applicable for finger-based haptic interaction (with haptic devices). This hand tracking module partially solves the magnetic-interferecne issues only for haptic device, and loses its generality when interacting with other magnetic objects (steel-wall, electronics, powered-tool), thus, we consequently develop the first tightly-coupled visual-inertial tracking framework with its experimental evaluation(chapter 4), which is much more accurate, robust, and generalizable for any challenging hand tracking application.

In the chapter 2, the detection threshold of finger tracking error, which would be a significant design specification for every hand tracking system, is revealed from the human subject studies. From the mean and variance of the average correct answer rate of discriminating the true finger-tip, we can obtain the detection threshold of this visual-proprioceptive conflict (i.e., trackiin error). The obtained detection threshold value to be 5.11cm from the human subject study. This value would be meaningful for designing hand tracking system since every hand tracking should satisfy at least for the applications where the tracking error would not substantially degrade user perception.

Moreover, in case of delivering the realistic haptic feedback to the human subjects, as compared to the detection threshold 5.11cm for the case of no haptic feedback, the detection threshold is increased to This is quantitatively evidenced by their respective detection to 6.05cm. This finding manifest that suitable usage of haptic feedback in finger-based interaction would be desirable in terms of user experience and allowing for lower tracking accuracy of hand tracking system, which is the reason we construct the finger tracking module compatible with haptic device in the next chapter. Given the fact that any tracking system can never be perfect due to the limitation of the deployed sensors, (e.g., vision-based tracking systems or wearable-sensor-based systems), we believe the above findings would be very useful to design the performance specification of finger tracking system with haptic feedback, which. The exact numbers associated to our findings would depend on the specifics of the experiments and their setup, e.g., size of finger spheres, construction of the virtual environment, dynamic perturbation error, etc. Even so, we strongly believe that the (vivid) trends of our findings and also the framework to obtain those would still be relevant and equally useful to other haptically-enabled VR applications in practice as well.

Based on the above findings and intuition about hand tracking, in chapter 3 we introduce novel wearable finger tracking module, FTM, which can show the reasonable tracking performance compatible with haptic devices. The existing hand tracking techniques have fundamental limitations especially integrating with the additional extra devices, that is, occlusion and visual distortion for vision-based tracking systems, magnetic-interference for IMU/compass wearable systems, and complex arrangement/packing to estimate multi-DOFs joint for soft wearable systems. To address these issues, we utilize complementary wearable sensors (IMU/compass module and soft sensor) and deploy the sensors opportunistically considering the anatomical characteristic of human hands. Since our developed FTM is based on the heterogeneous sensors (i.e., soft sensors, IMUs), this sensor configuration allow for multi-DOF anatomically-consistent dexterous finger/hand motion tracking while avoiding motor-IMU magnetic interference in small-size form factor. The quantitative evaluation for tracking performance of each sensing modality is conducted by MOCAP system.

We then integrate the wearable FTM with the 3-DOF CHDs and construct the wearable cutaneous haptic interface, WCHI. The ped-in-hole task, which is difficult task successfully fulfilled in VR due to requirement of fair dexterity and precision, is implemented in VR and tested through the WHCI. We conduct human subject study to verify the performance/capability of the proposed FTM, and also desire to certify the efficacy of the integrated WCHI (i.e., wearable finger-based interface with ability to track finger motion precisely and deliver haptic feedback simultaneously) for real VR applications. In the experiments, the subjects report that when the abduction-adduction (AA) motion is precisely tracked (cf. which requires complex packing of soft sensors), the task is more achievable, which shows the validity of our FTM for real finger-based interaction. Moreover, the haptic feedback for the multi-fingered interaction is verified as a crucial factor to increase/enhance the user experience and task performance. This result clearly exhibits that both haptic feedback and accurate tracking are important for the finger-based manipulation task, which reinforces the necessity of accurate hand tracking.

However, the proposed FTM also solves the issues of existing method restrictively, since the system fundamentally employs the heterogeneous sensors (IMU/compass and soft sensor), not performs the sensor-fusion of the both sensors utilizing complementary aspects. To be more specific, the FTM well address the issues when attaching the CHDs, for the most problematic cases of each sensor, that is, interacting with magnetic objects which can be completely solved by eliminating the problematic compass, or mechanical contacts with objects of soft-sensor. The many erroneous factors of each sensing modality also remain such as drift problem (IMU), linear disturbance (IMU), limited ruggedness (soft sensor), dependency to external sensor due to no global positioning (both sensors), or sensor-attachment error (both sensors). Due to the existence of these issues, we aim to develop a novel complete hand tracking framework, which is immune to every stated issue and improve accuracy and robustness simultaneously, via the sensor fusion of complementary aspects, which is motivation for developing visual-inertial tracking system (VIST) for human hand.

In the chapter 4, we propose a novel framework, VIST, applicable to general skeleton tracking system, which has the superior performance (accuracy and robustness) and overcomes all the issues of existing systems. Representatively, the vision-based tracking systems and IMU/compass wearable tracking systems are widely used, thus, mainly compared with our VIST. The vision-based systems have fundamental issues of occlusion (self-occlusion, outside the FOV, object occlusion) and dataset-dependency, which refrain the applicability to the daily activities (where our hands are interacting with a myriad of daily objects) and general users in various environments (where all users has all different hand shapes in all different backgrounds). The IMU/compass wearable systems not only have lower accuracy than vision-based systems (yet much better immunity to occlusion than vision-based systems) but also have fundamental limitations of magnetic interference owing to the high dependency on compass sensor, which largely decreases the usability of the system (requiring frequent calibrations) and range of usage only for magnetic-free object (not including steel/magnets/electronics) or environment (not in smart factory, near steel wall).

On the contrary, our proposed VIST is immune to all the issues above, which means the first hand tracking system suitable for any object/user/environment, even with much improved tracking accuracy and robustness. Through the chapter 4, the hardware setup and algorithm are thoroughly described to explain what is the technological/methodological contributions of our VIST and why our framework achieves the improved performance. The vision sensor and IMU sensor, which has the complementary aspects (i.e., sensitive to occlusion, fast motion but drift-free information (vision) and high frequency rate with erroneous drift/bias), thus already widely used for many other estimation algorithm (e.g., localization and mapping for robots (Qin *et al.* 2018; Hesch *et al.* 2014; Chen *et al.* 2019)) are integrated and fused effectively in our VIST framework.

The hardware setup of VIST, which employs multiple IMUs, many, passive and anonymous visual markers, and a stereo camera, is optimized solution for tracking the human hand, which has peculiarity of large-DOFs in small size. Through this setup, we can estimate large DOFs motions without problematic compass sensor, by acquiring rich visual information of hands via a large number of markers. The type of markers are adopted as a passive type (color fabric), which has merit of simple implementation without electronic installation (e.g., VIVE tracker, LED) or large form-factor (e.g., AR marker, feature-based marker), simultaneously maximizing the number of markers attached on the small-size human hand.

The sensor fusion of these numerous markers with inertial information from IMUs, yet, are much difficult since most of markers are frequently unobserved for skeletal systems, particularly for hands, since the markers are all anonymous and frequently occluded (self-occlusion, object-occlusion) due to the mentioned peculiarity of human hands. Thus, we firstly propose the systemic framework of visual-inertial sensor fusion for skeletal system, which adopts TC-fusion framework. In the TC-fusion, where the vision and IMU module are tightly-coupled (i.e., coexistence of inertial-to-vision and vision-to-inertial loops), visual and inertial information can robustly be matched and optimally fused for enhanced hand tracking. Through the inertial-to-vision loop (i.e., IMU-aided correspondence search in Sec. 4.4), inertial information of IMUs are utilized to estimate observation probability of each marker, which facilitates maximum-likelihood matching of many anonymous markers even in case of occlusions or outside the

FOV. Via the visual-to-inertial loop (i.e., visual information based correction and filtering in Sec. 4.5), visual information, which is not anonymous but matched with the hand from above correspondence search, are utilized to estimate a large number of hand skeletons with out compass information, and even the erroneous kinematic-related parameter (i.e., hand scale, sensor-attachment error, IMU sensor biases) automatically real-time, which mainly degrade tracking accuracy and usability of IMU/compass (or even soft-sensor) wearable systems.

Through the experiments in the chapter 4.6 and chapter 4.8 with their mean error and PCK measures given in the figures, our VIST framework significantly outperforms other state-of-the-art vision-based systems (Mueller *et al.* 2018; Zimmermann & Brox 2017; Iqbal *et al.* 2018; Zhang *et al.* 2020), particular for some challenging scenarios (e.g., visually complex backgrounds/objects), for which existing vision-based methods even cannot work stably, whereas the VIST framework still retains its tracking performance. The successful tracking with CHDs (Lee *et al.* 2019) also manifest that our VIST framework can be easily ported for applications requiring extra wearable devices/attachments (e.g., robotic hand teleoperation with CHDs, soft prosthesis (Kang *et al.* 2019)).

We also show stable performance of the VIST framework for different types of occlusions and magnetic interference. Specifically, motion tracking while interacting with diverse objects has been an issue for existing hand tracking systems: vision-based systems not robust for untrained objects (Armagan *et al.* 2020) or severe object-occlusion (Hampali *et al.* 2020; Mueller *et al.* 2017); soft-sensor wearable systems susceptible to object mechanical contact (Park *et al.* 2017; Kim *et al.* 2016); IMU/compass or magnetic wearable systems (Lee *et al.* 2019; Baldi *et al.* 2017; Ma *et al.*

2011) fragile to objects with magnets or internal currents. Given that human hands ceaselessly interact with various objects in daily life, the robust hand tracking of our VIST framework implies its applicability for a variety of real-world applications with diverse objects, that have defied other approaches so far (e.g., daily activity monitoring for rehabilitation, skill training/assessment of equipment/tools).

In addition, we verify that the VIST system can robustly track hand motion outdoors, which is tough for most existing systems, as the sunlight interferes with many types of IR sensors (e.g., RGB-D camera (Sridhar *et al.* 2015; Mueller *et al.* 2017), external IR tracker required for wearable tracking systems (Lee *et al.* 2019; Baldi *et al.* 2017)), whereas outdoor hand-tracking datasets for machine learning are extremely 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).

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

References

Abbas, Syed Muhammad & Muhammad, Abubakr 2012 Outdoor rgb-d slam performance in slow mine detection. In *Proc. German Conference on Robotics*, pp. 1–6.

Allin, Sonya, Matsuoka, Yoky & Klatzky, Roberta 2002 Measuring just noticeable differences for haptic force feedback: implications for rehabilitation. In *Proc. IEEE Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems*, pp. 299–302.

Andrychowicz, Marcin, Baker, Bowen, Chociej, Maciek, Józefowicz, Rafal, McGrew, Bob, Pachocki, Jakub, Petron, Arthur, Plappert, Matthias, Powell, Glenn, Ray, Alex, Schneider, Jonas, Sidor, Szymon, Tobin, Josh, Welinder, Peter, Weng, Lilian & Zaremba, Wojciech 2020 Learning dexterous in-hand manipulation. *International Journal of Robotics Research* **39** (1), 3–20.

Armagan, Anil, Garcia-Hernando, Guillermo, Baek, Seungryul, Hampali, Shreyas, Rad, Mahdi, Zhang, Zhaohui, Xie, Shipeng, Chen, MingXiu, Zhang, Boshen, Xiong, Fu, Xiao, Yang, Cao, Zhiguo, Yuan, Junsong, Ren, Pengfei, Huang, Weiting, Sun, Haifeng, Hrúz, Marek, Kanis, Jakub, Krňoul, Zdeněk, Wan, Qingfu, Li, Shile, Yang, Linlin, Lee, Dongheui, Yao, Angela, Zhou, Weiguo, Mei, Sijia, Liu, Yunhui, Spurr, Adrian, Iqbal, Umar, Molchanov, Pavlo, Weinzaepfel, Philippe, Brégier, Romain, Rogez, Gregory, Lepetit, Vincent & Kim, TaeKyun 2020

Measuring generalisation to unseen viewpoints, articulations, shapes and objects for 3d hand pose estimation under hand-object interaction. In *Proc. European Conference on Computer Vision*, pp. 85–101.

Baldi, Tommaso Lisini, Mohammadi, Mostafa, Scheggi, Stefano & Prattichizzo, Domenico 2015 Using inertial and magnetic sensors for hand tracking and rendering in wearable haptics. In *Proc. IEEE World Haptics Conference*, pp. 381–387.

Baldi, Tommaso Lisini, Scheggi, Stefano, Meli, Leonardo, Mohammadi, Mostafa & Prattichizzo, Domenico 2017 Gesto: A glove for enhanced sensing and touching based on inertial and magnetic sensors for hand tracking and cutaneous feedback. *IEEE Transactions on Human-Machine Systems* **47** (6), 1066–1076.

Besl, Paul J & McKay, Neil D 1992 Method for registration of 3-d shapes. In *Proc. Sensor Fusion IV: Control Paradigms and Data Structures*, pp. 586–606.

Bimbo, Joao, Pacchierotti, Claudio, Aggravi, Marco, Tsagarakis, Nikos & Prattichizzo, Domenico 2017 Teleoperation in cluttered environments using wearable haptic feedback. In *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3401–3408.

Bleser, Gabriele, Hendeby, Gustaf & Miezal, Markus 2011 Using egocentric vision to achieve robust inertial body tracking under magnetic disturbances. In *Proc. IEEE International Symposium on Mixed and Augmented Reality*, pp. 103–109.

Burns, Eric, Razzaque, Sharif, Panter, Abigail T, Whitton, Mary C, McCallus, Matthew R & Brooks, Frederick P 2005 The hand is slower than the eye: A quantitative exploration of visual dominance over proprioception. In *Proc. IEEE Virtual Reality Conference*, pp. 3–10.

Chan, Ting Kwok, Yu, Ying Kin, Kam, Ho Chuen & Wong, Kin Hong 2018 Robust hand gesture input using computer vision, inertial measurement unit (imu) and flex sensors. In *Proc. IEEE International Conference on Mechatronics, Robotics and Automation*, pp. 95–99.

Chang, Lillian Y & Pollard, Nancy S 2008 Method for determining kinematic parameters of the in vivo thumb carpometacarpal joint. *IEEE Transactions on Biomedical Engineering* **55** (7), 1897–1906.

Chen, Yiming, Zhang, Mingming, Hong, Dongsheng, Deng, Chengcheng & Li, Mingyang 2019 Perception system design for low-cost commercial ground robots: Sensor configurations, calibration, localization and mapping. In *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 6663–6670.

Chinello, Francesco, Malvezzi, Monica, Pacchierotti, Claudio & Prattichizzo, Domenico 2015 Design and development of a 3RRS wearable fingertip cutaneous device. In *Proc. IEEE/ASME International Conference on Advanced Intelligent Mechatronics*, pp. 293–298.

Chinello, Francesco, Pacchierotti, Claudio, Malvezzi, Monica & Prattichizzo, Domenico 2017 A three revolute-revolute-spherical wearable fingertip cutaneous device for stiffness rendering. *IEEE Transactions on Haptics* **11** (1), 39–50.

Chossat, Jean-Baptiste, Tao, Yiwei, Duchaine, Vincent & Park, Yong-Lae 2015 Wearable soft artificial skin for hand motion detection with embedded microfluidic strain sensing. In *Proc. IEEE International Conference on Robotics and Automation*, pp. 2568–2573.

Colman, Andrew M 2015 *A dictionary of psychology*. New york, NY: Oxford University Press.

Drewing, Knut & Ernst, Marc O 2006 Integration of force and position cues for shape perception through active touch. *Brain Research* **1078** (1), 92–100.

Faessler, Matthias, Fontana, Flavio, Forster, Christian, Mueggler, Elias, Pizzoli, Matia & Scaramuzza, Davide 2016 Autonomous, vision-based flight and live dense 3d mapping with a quadrotor micro aerial vehicle. *Journal of Field Robotics* **33** (4), 431–450.

Folegatti, A., de Vignemont, F., Pavani, F., Rossetti, Y. & Farné, A. 2009 Losing one's hand: visual-proprioceptive conflict affects touch perception. *PLoS One* **4** (9), e6920.

Fontana, Marco, Fabio, Salsedo, Marcheschi, Simone & Bergamasco, Massimo 2013 Haptic hand exoskeleton for precision grasp simulation. *Journal of Mechanisms and Robotics* **5** (4), 041014.

Forster, Christian, Carlone, Luca, Dellaert, Frank & Scaramuzza, Davide 2016 On-manifold preintegration for real-time visual–inertial odometry. *IEEE Transactions on Robotics* **33** (1), 1–21.

Frati, Valentino & Prattichizzo, Domenico 2011 Using kinect for hand tracking and rendering in wearable haptics. In *Proc. IEEE World Haptics Conference*, pp. 317–321.

Frisoli, Antonio, Rocchi, Fabrizio, Marcheschi, Simone, Dettori, Andrea, Salsedo, Fabio & Bergamasco, Massimo 2005 A new force-feedback arm exoskeleton for haptic interaction in virtual environments. In *Proc. IEEE Joint Eurohaptics Conference and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems*, pp. 195–201.

Gao, Wei & Tedrake, Russ 2019 Filterreg: Robust and efficient probabilistic point-set registration using gaussian filter and twist parameterization. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11095–11104.

Glauser, Oliver, Wu, Shihao, Panozzo, Daniele, Hilliges, Otmar & Sorkine-Hornung, Olga 2019 Interactive hand pose estimation using a stretch-sensing soft glove. *ACM Transactions on Graphics* **38** (4), 1–15.

Gleeson, Brian T, Horschel, Scott K & Provancher, William R 2010 Design of a fingertip-mounted tactile display with tangential skin displacement feedback. *IEEE Transactions on Haptics* **3** (4), 279–301.

Hampali, Shreyas, Rad, Mahdi, Oberweger, Markus & Lepetit, Vincent 2020 Honnotate: A method for 3d annotation of hand and object poses. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3196–3206.

Hesch, Joel A, Kottas, Dimitrios G, Bowman, Sean L & Roumeliotis, Stergios I 2014 Camera-imu-based localization: Observability analysis and consistency improvement. *International Journal of Robotics Research* **33** (1), 182–201.

Hirose, Osamu 2020 A bayesian formulation of coherent point drift. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **Early Access**.

Hollister, Anne, Giurintano, David J, Buford, William L, Myers, Loyd M & Novick, Andrew 1995 The axes of rotation of the thumb interphalangeal and metacarpophalangeal joints. *Clinical Orthopaedics and Related Research* **320**, 188–193.

Holmes, Nicholas P & Spence, Charles 2005 Visual bias of unseen hand position with a mirror: spatial and temporal factors. *Experimental Brain Research* **166** (3-4), 489–497.

Hrabia, Christopher-Eyk, Wolf, Katrin & Wilhelm, Mathias 2013 Whole hand modeling using 8 wearable sensors: Biomechanics for hand pose prediction. In *Proc. Augmented Human International Conference*, pp. 21–28.

In, Hyunki, Kang, Brian Byunghyun, Sin, Minki & Cho, KyuJin 2015 Exo-glove: A wearable robot for the hand with a soft tendon routing system. *IEEE Robotics and Automation Magazine* **22** (1), 97–105.

Iqbal, Umar, Molchanov, Pavlo, Gall, Thomas Breuel Juergen & Kautz, Jan 2018 Hand pose estimation via latent 2.5 d heatmap regression. In *Proc. European Conference on Computer Vision*, pp. 118–134.

Jang, Inyoung & Lee, Dongjun 2014 On utilizing pseudo-haptics for cutaneous fingertip haptic device. In *Proc. IEEE Haptics Symposium*, pp. 635–639.

Kang, Brian Byunghyun, Choi, Hyungmin, Lee, Haemin & Cho, Kyu-Jin 2019 Exo-glove poly ii: A polymer-based soft wearable robot for the hand with a tendon-driven actuation system. *Soft Robotics* **6** (2), 214–227.

Kim, Daekyum, Kang, Brian Byunghyun, Kim, Kyu Bum, Choi, Hyungmin, Ha, Jeesoo, Cho, Kyu-Jin & Jo, Sungho 2019 Eyes are faster than hands: A soft wearable robot learns user intention from the egocentric view. *Science Robotics* **4** (26), eaav2949.

Kim, Dong Hyun, Lee, Sang Wook & Park, Hyung-Soon 2016 Improving kinematic accuracy of soft wearable data gloves by optimizing sensor locations. *Sensors* **16** (6), 766.

Kim, Jun-Sik & Park, Jung-Min 2015 Physics-based hand interaction with virtual objects. In *Proc. IEEE International Conference on Robotics and Automation*, pp. 3814–3819.

Kim, Keehoon, Youm, Youngil & Chung, Wan Kyun 2002 Human kinematic factor for haptic manipulation: The wrist to thumb. In *Proc. Haptics Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems*, pp. 319–326.

Kim, Myungsin, Lee, Yongjun, Lee, Yongseok & Lee, Dongjun 2017 Haptic rendering and interactive simulation using passive midpoint integration. *International Journal of Robotics Research* **36** (12), 1341–1362.

Kortier, Henk G, Sluiter, Victor I, Roetenberg, Daniel & Veltink, Peter H 2014 Assessment of hand kinematics using inertial and magnetic sensors. *Journal of neuroengineering and rehabilitation* **11** (1), 1–15.

Kuchenbecker, Katherine J, Ferguson, David, Kutzer, Michael, Moses, Matthew & Okamura, Allison M 2008 The touch thimble: Providing fingertip contact feedback during point-force haptic interaction. In *Proc. IEEE Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems*, pp. 239–246.

Lee, Jung Keun, Park, Edward J & Robinovitch, Stephen N 2012 Estimation of attitude and external acceleration using inertial sensor measurement during various dynamic conditions. *IEEE Transactions on Instrumentation and Measurement* **61** (8), 2262–2273.

Lee, Yongjun, Kim, Myungsin, Lee, Yongseok, Kwon, Junghan, Park, YongLae & Lee, D. J. 2019 Wearable finger tracking and cutaneous haptic interface with soft sensors for multi-fingered virtual manipulation. *IEEE/ASME Transactions on Mechatronics* **24** (1), 67–77.

Leonardis, Daniele, Solazzi, Massimiliano, Bortone, Ilaria & Frisoli, Antonio 2015 A wearable fingertip haptic device with 3 DoF asymmetric 3-RSR kinematics. In *Proc. IEEE World Haptics Conference*, pp. 388–393.

Li, Yue, Weng, Dongdong, Li, Dong & Wang, Yihan 2019 A low-cost drift-free optical-inertial hybrid motion capture system for high-precision human pose detection. In *Proc. IEEE International Symposium on Mixed and Augmented Reality (Adjunct)*, pp. 75–80.

Livingston, Mark A & Ai, Zhuming 2008 The effect of registration error on tracking distant augmented objects. In *Proc. IEEE International Symposium on Mixed and Augmented Reality*, pp. 77–86.

Luinge, Henk J, Veltink, Peter H & Baten, Chris TM 2007 Ambulatory measurement of arm orientation. *Journal of Biomechanics* **40** (1), 78–85.

Lynen, Simon, Achtelik, Markus W, Weiss, Stephan, Chli, Margarita & Siegwart, Roland 2013 A robust and modular multi-sensor fusion approach applied to mav navigation. In *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3923–3929.

Ma, Yinghong, Mao, Zhi-Hong, Jia, Wenyan, Li, Chengliu, Yang, Jiawei & Sun, Mingui 2011 Magnetic hand tracking for human-computer interface. *IEEE Transactions on Magnetics* **47** (5), 970–973.

Madsen, Jacob B & Stenholt, Rasmus 2014 How wrong can you be: Perception of static orientation errors in mixed reality. In *Proc. IEEE Symposium on 3D User Interfaces*, pp. 83–90.

Maereg, Andualem Tadesse, Nagar, Atulya, Reid, David & Secco, Emanuele L 2017 Wearable vibrotactile haptic device for stiffness discrimination during virtual interactions. *Frontiers in Robotics and AI* **4**, 42.

Mahony, Robert, Hamel, Tarek & Pflimlin, Jean-Michel 2008 Nonlinear complementary filters on the special orthogonal group. *IEEE Transactions on Automatic Control* **53** (5), 1203–1218.

Maisto, Maurizio, Pacchierotti, Claudio, Chinello, Francesco, Salvietti, Gionata, De Luca, Alessandro & Prattichizzo, Domenico 2017 Evaluation of wearable haptic systems for the fingers in augmented reality applications. *IEEE Transactions on Haptics* **10** (4), 511–522.

Mallat, Randa, Bonnet, Vincent, Khalil, Mohamad Ali & Mohammed, Samer 2020 Upper limbs kinematics estimation using affordable visual-inertial sensors. *IEEE Transactions on Automation Science and Engineering* **Early Access**.

Markley, F Landis 1988 Attitude determination using vector observations and the singular value decomposition. *The Journal of the Astronautical Sciences* **36** (3), 245–258.

Meilland, Maxime, Comport, Andrew I & Rives, Patrick 2015 Dense omnidirectional rgb-d mapping of large-scale outdoor environments for real-time localization and autonomous navigation. *Journal of Field Robotics* **32** (4), 474–503.

Meli, Leonardo, Pacchierotti, Claudio, Salvietti, Gionata, Chinello, Francesco, Maisto, Maurizio, De Luca, Alessandro & Prattichizzo, Domenico 2018 Combining wearable finger haptics and augmented reality: User evaluation using an external camera and the microsoft hololens. *IEEE Robotics and Automation Letters* **3** (4), 4297–4304.

Meli, Leonardo, Scheggi, Stefano, Pacchierotti, Claudio & Prattichizzo, Domenico 2013 Towards wearability in fingertip haptics: a 3-dof wearable device for cutaneous force feedback. *IEEE Transactions on Haptics* **6** (4), 506–516.

Meli, Leonardo, Scheggi, Stefano, Pacchierotti, Claudio & Prattichizzo, Domenico 2014 Wearable haptics and hand tracking via an rgb-d camera for immersive tactile experiences. In *Proc. ACM SIGGRAPH Posters*, p. 56.

Minamizawa, Kouta, Fukamachi, Souichiro, Kajimoto, Hiroyuki, Kawakami, Naoki & Tachi, Susumu 2007 Gravity grabber: wearable haptic display to present virtual mass sensation. In *Proc. ACM SIGGRAPH Emerging Technologies*, p. 8.

Mizera, C, Delrieu, T, Weistroffer, V, Andriot, C, Decatoire, A & Gazeau, J-P 2019 Evaluation of hand-tracking systems in teleoperation and virtual dexterous manipulation. *IEEE Sensors Journal* **20** (3), 1642–1655.

Moon, Gyeongsik, Ju, YongChang & Lee, KyoungMu 2018 V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5079–5088.

Mueller, Franziska, Bernard, Florian, Sotnychenko, Oleksandr, Mehta, Dushyant, Sridhar, Srinath, Casas, Dan & Theobalt, Christian 2018 Ganerated hands for real-time 3d hand tracking from monocular rgb. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 49–59.

Mueller, Franziska, Mehta, Dushyant, Sotnychenko, Oleksandr, Sridhar, Srinath, Casas, Dan & Theobalt, Christian 2017 Real-time hand tracking under occlusion from an egocentric rgb-d sensor. In *Proc. IEEE International Conference on Computer Vision*, pp. 1284–1293.

Muth, Joseph T, Vogt, Daniel M, Truby, Ryan L, Mengüç, Yiğit, Kolesky, David B, Wood, Robert J & Lewis, Jennifer A 2014 Embedded 3d printing of strain sensors within highly stretchable elastomers. *Advanced Materials* **26** (36), 6307–6312.

Myronenko, Andriy & Song, Xubo 2010 Point set registration: Coherent point drift. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **32** (12), 2262–2275.

Oikonomidis, Iason, Kyriazis, Nikolaos & Argyros, Antonis A 2011 Efficient model-based 3d tracking of hand articulations using kinect. In *Proc. British Machine Vision Conference*, , vol. 1, p. 3.

Pacchierotti, Claudio, Chinello, Francesco, Malvezzi, Monica, Meli, Leonardo & Prattichizzo, Domenico 2012 Two finger grasping simulation with cutaneous and kinesthetic force feedback. In *Proc. International Conference on Human Haptic Sensing and Touch Enabled Computer Applications*, pp. 373–382.

Panteleris, Paschalis & Argyros, Antonis 2017 Back to rgb: 3d tracking of hands and hand-object interactions based on short-baseline stereo. In *Proc. IEEE International Conference on Computer Vision (Workshops)*.

Park, Wookeun, Ro, Kyongkwan, Kim, Suin & Bae, Joonbum 2017 A soft sensor-based three-dimensional (3-d) finger motion measurement system. *Sensors* **17** (2), 420.

Pavlovic, Vladimir I, Sharma, Rajeev & Huang, Thomas S. 1997 Visual interpretation of hand gestures for human-computer interaction: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **19** (7), 677–695.

Perez, Alvaro G, Cirio, Gabriel, Lobo, Daniel, Chinello, Francesco, Prattichizzo, Domenico & Otaduy, Miguel A 2016 Efficient nonlinear skin simulation for multi-finger tactile rendering. In *Proc. IEEE Haptics Symposium*, pp. 155–160.

Peters, Michael, Mackenzie, Kevin & Bryden, Pam 2002 Finger length and distal finger extent patterns in humans. *American Journal of Physical Anthropology* **117** (3), 209–217.

Prachyabrued, Mores & Borst, Christoph W 2015 Design and evaluation of visual interpenetration cues in virtual grasping. *IEEE Transactions on Visualization and Computer Graphics* **22** (6), 1718–1731.

Prattichizzo, Domenico, Pacchierotti, Claudio, Cenci, Stefano, Minamizawa, Kouta & Rosati, Giulio 2010*a* Using a fingertip tactile device to substitute kinesthetic feedback in haptic interaction. In *Haptics: Generating and Perceiving Tangible Sensations*, pp. 125–130. Springer.

Prattichizzo, Domenico, Pacchierotti, Claudio, Cenci, Stefano, Minamizawa, Kouta & Rosati, Giulio 2010*b* Using a fingertip tactile device to substitute kinesthetic feedback in haptic interaction. In *Proc. International Conference on Human Haptic Sensing and Touch Enabled Computer Applications*, pp. 125–130.

Prattichizzo, Domenico, Pacchierotti, Claudio & Rosati, Giulio 2012*a* Cutaneous force feedback as a sensory subtraction technique in haptics. *IEEE Transactions on Haptics* **5** (4), 289–300.

Prattichizzo, Domenico, Pacchierotti, Claudio & Rosati, Giulio 2012*b* Cutaneous force feedback as a sensory subtraction technique in haptics. *IEEE Transactions on Haptics* **5** (4), 289–300.

Qin, Tong, Li, Peiliang & Shen, Shaojie 2018 Vins-mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Transactions on Robotics* **34** (4), 1004–1020.

Quek, Zhan Fan, Schorr, Samuel B, Nisky, Ilana, Provancher, William R & Okamura, Allison M 2015 Sensory substitution and augmentation using 3-degree-of-freedom skin deformation feedback. *IEEE Transactions on Haptics* **8** (2), 209–221.

Richard M. Murray, Zexiang Li, S. & Sastry, Shankar 1993 *A mathematical introduction to robotic manipulation*. Boca Ranton, FL: CRC press.

Roetenberg, Daniel, Luinge, Henk J, Baten, Chris TM & Veltink, Peter H 2005 Compensation of magnetic disturbances improves inertial and magnetic sensing of human body segment orientation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **13** (3), 395–405.

Santaera, Gaspare, Luberto, Emanuele, Serio, Alessandro, Gabiccini, Marco & Bicchi, Antonio 2015 Low-cost, fast and accurate reconstruction of robotic and human postures via imu measurements. In *Proc. IEEE International Conference on Robotics and Automation*, pp. 2728–2735.

Santello, Marco, Flanders, Martha & Soechting, John F 1998 Postural hand synergies for tool use. *Journal of Neuroscience* **18** (23), 10105–10115.

Scheggi, S., Meli, L., Pacchierotti, C. & Prattichizzo, D. 2015 Scheggi, stefano and meli, leonardo and pacchierotti, claudio and prattichizzo, domenico. In *Proc. ACM SIGGRAPH Posteres*, pp. 31–31.

Schorr, Samuel Benjamin & Okamura, Allison M 2017 Three-dimensional skin deformation as force substitution: Wearable device design and performance during haptic exploration of virtual environments. *IEEE Transactions on Haptics* **10** (3), 418–430.

Seel, Thomas, Raisch, Jörg & Schauer, Thomas 2014 Imu-based joint angle measurement for gait analysis. *Sensors* **14** (4), 6891–6909.

Shen, Shaojie, Mulgaonkar, Yash, Michael, Nathan & Kumar, Vijay 2013 Vision-based state estimation for autonomous rotorcraft mavs in complex environments. In *Proc. IEEE International Conference on Robotics and Automation*, pp. 1758–1764.

Solazzi, Massimiliano, Frisoli, Antonio & Bergamasco, Massimo 2010 Design of a cutaneous fingertip display for improving haptic exploration of virtual objects. In *Proc. IEEE International Symposium on Robot and Human Interactive Communication*, pp. 1–6.

Sridhar, Srinath, Mueller, Franziska, Oulasvirta, Antti & Theobalt, Christian 2015 Fast and robust hand tracking using detection-guided optimization. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3213–3221.

Sridhar, Srinath, Mueller, Franziska, Zollhöfer, Michael, Casas, Dan, Oulasvirta, Antti & Theobalt, Christian 2016 Real-time joint tracking of a hand manipulating an object from rgb-d input. In *Proc. European Conference on Computer Vision*, pp. 294–310.

Tan, Hong Z, Srinivasan, Mandayam A, Reed, Charlotte M & Durlach, Nathaniel I 2007 Discrimination and identification of finger joint-angle position using active motion. *ACM Transactions on Applied Perception* **4** (2), 10–es.

Tao, Yaqin, Hu, Huosheng & Zhou, Huiyu 2007 Integration of vision and inertial sensors for 3d arm motion tracking in home-based rehabilitation. *International Journal of Robotics Research* **26** (6), 607–624.

Tkach, Anastasia, Tagliasacchi, Andrea, Remelli, Edoardo, Pauly, Mark & Fitzgibbon, Andrew 2017 Online generative model personalization for hand tracking. *ACM Transactions on Graphics* **36** (6), 1–11.

Tobergte, Andreas, Helmer, Patrick, Hagn, Ulrich, Rouiller, Patrice, Thielmann, Sophie, Grange, Sébastien, Albu-Schäffer, Alin, Conti, François & Hirzinger, Gerd 2011 The sigma. 7 haptic interface for mirosurge: A new bi-manual surgical console. In *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3023–3030.

Tompson, Jonathan, Stein, Murphy, Lecun, Yann & Perlin, Ken 2014 Real-time continuous pose recovery of human hands using convolutional networks. *ACM Transactions on Graphics* **33** (5), 1–10.

Trawny, Nikolas & Roumeliotis, Stergios I 2005 Indirect kalman filter for 3d attitude estimation. *Techical Report, University of Minnesota* **2**.

Trindade, Pedro, Lobo, Jorge & Barreto, Joao P 2012 Hand gesture recognition using color and depth images enhanced with hand angular pose data. In *Proc. IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, pp. 71–76.

Van Beers, Robert J, Sittig, Anne C & Van Der Gon, Jan J Denier 1998 The precision of proprioceptive position sense. *Experimental Brain Research* **122** (4), 367–377.

Wang, Dangxiao, Song, Meng, Naqash, Afzal, Zheng, Yukai, Xu, Weiliang & Zhang, Yuru 2018 Toward whole-hand kinesthetic feedback: A survey of force feedback gloves. *IEEE Transactions on Haptics* **12** (2), 189–204.

Weber, Paul, Rueckert, Elmar, Calandra, Roberto, Peters, Jan & Beckerle, Philipp 2016 A low-cost sensor glove with vibrotactile feedback and multiple finger joint and hand motion sensing for human-robot interaction. In *Proc. IEEE International Symposium on Robot and Human Interactive Communication*, pp. 99–104.

Weiss, Stephan M 2012 Vision based navigation for micro helicopters. PhD thesis, ETH Zurich.

Welch, Greg & Foxlin, Eric 2002 Motion tracking survey: No silver bullet, but a respectable arsenal. *IEEE Computer Graphics and Appilications* **22** (6), 24–38.

Wichmann, Felix A & Hill, N Jeremy 2001 The psychometric function: I. fitting, sampling, and goodness of fit. *Perception and psychophysics* **63** (8), 1293–1313.

Wong, Charence, Zhang, Zhi-Qiang, Lo, Benny & Yang, Guang-Zhong 2015 Wearable sensing for solid biomechanics: A review. *IEEE Sensors Journal* **15** (5), 2747–2760.

Yang, Yi & Ramanan, Deva 2012 Articulated human detection with flexible mixtures of parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **35** (12), 2878–2890.

Yuan, Qilong, Chen, I-Ming & Sin, Ang Wei 2013 Method to calibrate the skeleton model using orientation sensors. In *Proc. IEEE International Conference on Robotics and Automation*, pp. 5297–5302.

Yuan, Shanxin, Ye, Qi, Stenger, Bjorn, Jain, Siddhant & Kim, Tae-Kyun 2017 Bighand2. 2m benchmark: Hand pose dataset and state of the art analysis. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4866–4874.

Zhang, Fan, Bazarevsky, Valentin, Vakunov, Andrey, Tkachenka, Andrei, Sung, George, Chang, Chuo-Ling & Grundmann, Matthias 2019*a* Mediapipe hands: On-device real-time hand tracking. In *Proc. IEEE International Conference on Computer Vision (Workshops).*

Zhang, Jiawei, Jiao, Jianbo, Chen, Mingliang, Qu, Liangqiong, Xu, Xiaobin & Yang, Qingxiong 2017 A hand pose tracking benchmark from stereo matching. In *Proc. IEEE International Conference on Image Processing*, pp. 982–986.

Zhang, Mingming, Chen, Yiming & Li, Mingyang 2019*b* Vision-aided localization for ground robots. In *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2455–2461.

Zhang, Zhaohui, Xie, Shipeng, Chen, Mingxiu & Zhu, Haichao 2020 Handaugment: A simple data augmentation method for depth-based 3d hand pose estimation. *arXiv preprint arXiv:2001.00702* .

Zhou, Shengli, Fei, Fei, Zhang, Guanglie, Mai, John D, Liu, Yunhui, Liou, Jay YJ & Li, Wen J 2013 2d human gesture tracking and recognition by the fusion of mems inertial and vision sensors. *IEEE Sensors Journal* **14** (4), 1160–1170.

Zimmermann, Christian & Brox, Thomas 2017 Learning to estimate 3d hand pose from single rgb images. In *Proc. IEEE International Conference on Computer Vision*, pp. 4903–4911.

인간 기계 상호작용을 위한 강건하고 정확한 손동작 추적 기술 연구

서울대학교 대학원
기계항공공학부
이 용 석

요 약

손 동작을 기반으로 한 인터페이스는 인간-기계 상호작용 분야에서 직관성, 몰입감, 정교함을 제공해줄 수 있어 많은 주목을 받고 있고, 이를 위해 가장 필수적인 기술 중 하나가 손 동작의 강건하고 정확한 추적 기술 이다. 이를 위해 본 학위논문에서는 먼저 사람 인지의 관점에서 손 동작 추적 오차의 인지 범위를 규명한다. 이 오차 인지 범위는 새로운 손 동작 추적 기술 개발 시 중요한 설계 기준이 될 수 있어 이를 피험자 실험을 통해 정량적으로 밝히고, 특히 손끝 촉각 장비가 있을때 이 인지 범위의 변화도 밝힌다. 이를 토대로, 촉각 피드백을 주는 것이 다양한 인간-기계 상호작용 분야에서 널리 연구되어 왔으므로, 먼저 손끝 촉각 장비와 함께 사용할 수 있는 손 동작 추적 모듈을 개발한다. 이 손끝 촉각 장비는 자기장 외란을 일으켜 착용형 기술에서 흔히 사용되는 지자기 센서를 교란하는데, 이를 적절한 사람 손의

해부학적 특성과 관성 센서/지자기 센서/소프트 센서의 적절한 활용을 통해 해결한다. 이를 확장하여 본 논문에서는, 촉각 장비 착용 시 뿐 아니라 모든 장비 착용 / 환경 / 물체와의 상호작용 시에도 사용 가능한 새로운 손 동작 추적 기술을 제안한다. 기존의 손 동작 추적 기술들은 가림 현상 (영상 기반 기술), 지자기 외란 (관성/지자기 센서 기반 기술), 물체와의 접촉 (소프트 센서 기반 기술) 등으로 인해 제한된 환경에서 밖에 사용하지 못한다. 이를 위해 많은 문제를 일으키는 지자기 센서 없이 상보적인 특성을 지니는 관성 센서와 영상 센서를 융합하고, 이때 작은 공간에 다 자유도의 움직임을 갖는 손 동작을 추적하기 위해 다수의 구분되지 않는 마커들을 사용한다. 이 마커의 구분 과정 (correspondence search)를 위해 기존의 약결합 (loosely-coupled) 기반이 아닌 강결합 (tightly-coupled 기반 센서 융합 기술을 제안하고, 이를 통해 지자기 센서 없이 정확한 손 동작이 가능할 뿐 아니라 착용형 센서들의 정확성/편의성에 문제를 일으키던 센서 부착 오차 / 사용자의 손 모양 등을 자동으로 정확히 보정한다. 이 제안된 영상-관성 센서 융합 기술 (Visual-Inertial Skeleton Tracking (VIST)) 의 뛰어난 성능과 강건성이 다양한 정량/정성 실험을 통해 검증되었고, 이는 VIST의 다양한 일상환경에서 기존 시스템이 구현하지 못하던 손 동작 추적을 가능케 함으로써, 많은 인간-기계 상호작용 분야에서의 가능성을 보여준다.

주요어: 손동작 추적, 강결합 센서 융합, 비선형 칼만 필터링, 정보 일치 탐색, 관성 센서, 컴퓨터 비전, 인간-컴퓨터 상호작용, 인간-로봇 상호작용, 가상 현실, 증강 현실, 촉각 피드백

학 번: 2013-23082

Acknowledgement

장장 12년, 제 인생의 절반 가까이 되는 시간을 보낸 관악에서의 생활이 끝이 났습니다. 돌이켜보면 제 삶의 여러 선택의 기로마다 저는, 스스로가 가장 성장할 수 있는 길이 무엇인지를 고민하고 선택해온 것 같습니다. 비록 이러한 선택들에 대해 후회하기도 최선을 다하지 못한 자신에 대해 질책하기도 하였지만, 그럼에도 오늘만큼은 스스로에게 "그동안 정말 애썼다. 지금까지 노력해줘서 정말로 고맙다!"고 말하고 싶습니다.

제가 학위 과정을 무사히 마치게 되기까지 수많은 사람들의 노력이 있었음을 알기에 감사의 글을 남기고자 합니다. 무엇보다 제 지도교수님이신 이동준 교수님께 너무나 감사드립니다. 교수님께서 몸소 보여주신 열정과 헌신, 창의성은 저에게 많은 자극과 영감을 주었고, 교수님을 본받기 위해 노력하다 보니 이만큼이나 성장 할 수 있다고 생각합니다. 학위 과정 동안 교수님께 배운 지식과 삶의 태도 모두 소중히 간직하여 끊임없이 발전하는 연구자가 될 수 있도록 하겠습니다.

또한 제 가족들, 지금의 저를 있게 해주신 부모님, 무한한 사랑으로 저를 길러주신 할머니, 부족한 형보다 더 형처럼 듬직한 모습을 보여준 동생 형돈이와 저를 자식처럼 아껴주시고 격려해주신 이모님들께 정말 감사드립니다. 특히 지난 30년간 단 한 번도 저를 의심하지 않고 한결같은 믿음과 희생을 보여주신 어머니께 정말 감사하다고, 어머니의 아들이어서 정말 행복하다고 말씀드리고 싶습니다.

연구 내적/외적으로 많은 도움을 준 연구실 사람들에게도 감사를 전합니다. 8년 동안 개인적으로도 연구적으로도 많은 의지가 되어준 현수, 신입생 때부터 연구 방향을 잡는 데에 많은 조언을 주셨던 명신이형과 창수형, 함께 연구를 진행하며 많은 도움을 받았던 인영, 용준, 원하와 연구 과제로 드론을 날리며 수많은 고생을 함께 한 용한, 지석, 재민, 정섭이, 그리고 선배로서 지도는커녕 졸업한다는 핑계로 궂은일만 시켰음에도 자기 일처럼 도와준 진욱, 소망, 영선이를 비롯, 저와 함께한 모든 연구실 동료들에게 정말로 감사드립니다. 또한, 오랜 시간 함께하며 제 연구에 크게 기여했을 뿐 아니라, 제가 어려운 시간을 이겨낼 수 있게 도와준 원경이와 연구 진행에 큰 도움을 준 한별이, 병철이에게도 감사의 인사를 전합니다.

마지막으로 지금까지 제 가장 힘들었던 날들과 가장 즐거운 날들을 모두 함께한 제 소중한 친구들인 두기, 중경, 민규, 아민, 건호, 세현이에게 함께 해서 정말로 즐거웠고, 그래서 정말로 고마웠다고 전하고 싶습니다. 또한, 오랜 시간 우정을 쌓아온 제 고등학교 친구들인 현택, 윤혁, 덕수, 용민, 택규와 만날 때마다 격려해주고 응원해준 기계과 동기들인 인욱, 근진, 동욱, 원준, 도현이 그리고 전공 수업의 조모임으로써 시작하여 소중한 인연을 이어오고 있는 광환이형, 준민, 승헌이형, 성희누나, 석원이형에게도 그동안 고마웠고 앞으로도 계속 소중한 인연 함께 이어가자고 말하고 싶습니다.

비록 여기에 다 적지는 못하였지만, 이 밖에도 수많은 사람들에게 큰 도움을 받으며 살아왔습니다. 저의 성취들이 단지 저만의 노력으로 이뤄진 것이 아님을 알기에 평생을 감사하는 마음으로, 또 제가 받은 것 이상을 베풀며 살도록 노력하겠습니다. 모두 감사합니다!