Journal of Intelligent Systems and Control
https://www.acadlore.com/journals/JISC

Bayesian Estimation of Hand Kinematics from Spatially Tracked Landmarks

Yiyang Dong, Shahram Payandeh*

School of Engineering Science, Simon Fraser University, V5A 1S6 Burnaby, Canada

* Correspondence: Shahram Payandeh (payandeh@sfu.ca)

Received: 03-10-2025 **Revised:** 05-28-2025 **Accepted:** 06-09-2025

Citation: Y. Y. Dong and S. Payandeh, "Bayesian estimation of hand kinematics from spatially tracked landmarks," *J. Intell Syst. Control*, vol. 4, no. 2, pp. 105–124, 2025. https://doi.org/10.56578/jisc040203.

Abstract: A Bayesian framework for estimating finger joint kinematics from spatially tracked hand landmarks was introduced in this study. Three-dimensional landmark data were constructed by augmenting image-based two-dimensional hand landmarks with calibrated depth information. A hierarchy of coordinate frames was established, beginning with the palm as the root and extending to child frames assigned to each finger, thereby encoding the natural kinematic dependencies of the hand. This hierarchical representation provides the structural foundation for Bayesian estimation. Finger joint parameters were estimated within a maximum likelihood framework that is robust to tracking noise and signal occlusions, which are common in practical hand-tracking scenarios. Unlike data-driven methods, the proposed approach does not rely on pre-collected training datasets but instead leverages the kinematic model and intrinsic physical constraints of the human hand. The estimation problem was formalized as a Gaussian Bayesian Network (GBN), through which joint parameters were inferred using Maximum Likelihood Estimation (MLE). Robustness of the approach was qualitatively demonstrated through reconstructed graphical configurations that illustrate accurate recovery of finger postures under noisy conditions. This method provides a principled framework for hand motion reconstruction and establishes the foundation for future quantitative evaluations against benchmark datasets. The framework is expected to advance applications in human–computer interaction, prosthetic design, virtual reality (VR), and rehabilitation by enabling more reliable and anatomically consistent hand tracking.

Keywords: Hand kinematic; Spatial landmarks; Bayesian model estimation; Maximum likelihood method; Configuration reconstruction

1 Introduction

Tracking and kinematic reconstruction of the human hand is an area of study with many potential applications such as robotics, gaming, or human-machine interface [1–5]. For example, in clinical settings, such as rehabilitation centers, accurate tracking of hand or arm movements and their reconstruction in virtual environments enable detailed data collection and visualization, essential for healthcare providers monitoring recovery in patients like injured athletes or elderly individuals. Advanced tracking and reconstruction algorithms, combined with state-of-the-art hardware, can offer clinicians deeper insights into a patient's motor abilities, providing a foundation for more accurate recovery assessments and aiding in early diagnosis of neurodegenerative conditions, including stroke [6], Alzheimer's disease [7] or amyotrophic lateral sclerosis (ALS) [8]. Hand tracking [5, 9–11] and kinematic reconstruction still face many challenges due to the high-dimensional configuration space of the hand, the motion and appearance variations, the limitations of cameras, and the interference of cluttered environmental backgrounds.

Designing effective hand tracking and reconstruction systems presents unique challenges. The human hand's complex structure, with its high degrees of freedom (DOF), makes it difficult to accurately track each joint and bone movement, as it involves estimating numerous parameters. Additional challenges arise from self-occlusions, where parts of the hand obscure others, and from the rapid speed of hand movements. These systems must also process substantial data volumes in real time, remaining robust in uncontrolled environments with noisy backgrounds and varying lighting conditions, all of which demand significant computational power and advanced algorithms.

Methods of hand tracking can be divided into two categories: appearance- and model-based approaches [1]. Methods based on visual appearance try to find a mapping from the input data to the hand pose space by utilizing features such as edges. These methods are computationally less demanding since the learned tracking information

is encoded for direct association without searching for the whole configuration space of the hand. However, these methods are limited by noise and partial occlusion in the input data and suffer from low accuracy and poor stability. Wang and Payandeh [12] used a Kalman filter to track human hands and leverage scale-invariant feature transform to extract features for hand posture recognition. Ge et al. [13] designed Multi-view Convolutional Neural Networks (MVCNNs) to detect hand pose from a single depth image. Its main idea is projecting the depth image onto three orthogonal planes and using these projections to regress 2D joint heat maps that are further fused to predict 3D hand pose. Zhang et al. [14] presented MediaPipe Hands that can achieve real-time hand tracking using a neural network. It first used a palm detector to predict a bounding box and then used a landmark model to predict the hand skeleton. A feature extractor was trained with extensive hand data, and it can output the 2.5D position of 21 landmarks, the handedness, and the probability of hand presence.

Model-based methods try to fit a hand model to the input data. This process is always done through optimization or nearest neighborhood search. The optimization-based methods try to fine-tune the model parameters of the 3D hand until the model fits the input data well. The methods based on nearest neighborhood search try to find the most similar hand pose in a big pose database. Oikonomidis et al. [15] presented a tracking method that can predict the position, orientation, and full articulation of a hand by using a Kinect sensor. It is a typically model-based method as it formulates this as an optimization problem and seeks for model parameters that minimize the discrepancy between the hand model and hand observation. To solve this problem, Particle Swarm Optimization (PSO) was used. Tagliasacchi et al. [16] presented a robust Articulated Iterative Closest Point (articulated-ICP) to achieve real-time hand tracking using a single depth camera. The articulated registration algorithm can integrate data and regularization priors into a unified solver to enable robust tracking without restricting the freedom of motion. Qian et al. [17] presented a hand tracking system in real time on a desktop. This method follows a local optimization and initialization by a part detection framework. A simple hand model approximated by a set of spheres was used and a cost function measuring the distance between the model and a sparse point cloud was designed. Moreover, some works have tried to use multimodal sensors to capture the motion of human hands. Diaz and Payandeh [18] investigated the integration of a multimodal sensing system for exploring limits of vibrato tactile haptic feedback when interacting with 3D representations of objects.

Several studies have developed kinematic models for robotic hands, each designed to address specific functional needs and structural requirements. For example, Tarmizi et al. [19] presented a multi-fingered robotic hand model, developed using the Denavit-Hartenberg algorithm and Euler-Lagrange equations, to capture joint configurations and movement dynamics accurately. Similarly, Lee et al. [20] introduced an anthropomorphic robotic hand, the Korea Institute of Industrial Technology Hand (KITECH-Hand), with a detailed kinematic analysis. This model uniquely focuses on the metacarpophalangeal (MCP) joints of the fingers, implementing a roll-pitch configuration to replace conventional yaw-pitch structures for improved mechanical functionality. Hand pose estimation methods leverage deep learning models, including diffusion-based approaches [21–24]. For instance, Wang et al. [21] modeled hand pose forecasting as a reverse diffusion process, introducing a dual-diffusion mechanism that simultaneously captures local and global hand features using specialized global and local diffusion blocks. Cheng et al. [22] proposed a diffusion-based model that de-noises hand pose predictions by integrating image and point cloud data. Joint-wise and local detail conditions were introduced to recover precise key-point positions. Additionally, Ye et al. [24] focused on hand-object interactions, designing a diffusion network that models the conditional distribution of object geometries based on hand configurations captured from video sequences.

As a more recent comparative literature review, Mangalam et al. [25] targeted realistic and immersive hand-object interaction in virtual reality (VR). Current limitations in VR systems were identified, such as unnatural post-collision behavior and ineffective grasping, and a multi-pronged solution across hand modeling, contact dynamics, and grasp release was proposed. The focus was on enhancing realism through physical simulation, neuroscientific insights, and mesh-based modeling, rather than explicit kinematic parameter estimation which is the main focus of this current study. While both works aim to improve hand representation in digital environments, the contributions of this current study emphasize kinematic inference and model estimation from data, whereas the referenced study addresses interaction realism and dynamic behavior in virtual environments. Joseph Isaac et al. [26] presented a deep learning-based approach that uses a convolutional neural network (CNN) trained on large, annotated datasets (HANDS2017 and MSRA) to directly predict 3D hand poses. Its novelty lies in enforcing anatomical correctness within the network architecture itself, avoiding the need for post-processing filters. The model ensures that predicted poses lie within the biomechanical constraints of the human hand, and even corrects errors found in the "ground truth" annotations by creating new datasets (AEF-HANDS2017 and AEF-MSRA). In contrast, this current study offers a data-efficient, model-driven solution that avoids training entirely, while the Semi-Supervised Cross-Correlation Convolutional Neural Network (SSC-CNN) offers a data-intensive, learning-based solution that corrects anatomical inconsistencies through architectural constraints. The former emphasizes hierarchical kinematic reasoning and interpretability, whereas the latter prioritizes learning anatomical validity within a predictive neural framework. Pavlakos et al. [27] presented a fully data-driven, transformer-based method for recovering full 3D hand meshes

from monocular images. It emphasizes scale and learning capacity, combining multiple annotated datasets and utilizing a large Vision Transformer (ViT) to achieve state-of-the-art accuracy. It excels in generalizing to in-the-wild conditions, as shown by performance gains on a new benchmark dataset (HInt) with real-world images and 2D keypoint annotations.

While this current study and the study by Pavlakos et al. [27] both aim to reconstruct the 3D structure of the hand, the former emphasizes interpretable, physically grounded kinematic estimation without the need for training data, whereas the latter focuses on high-capacity, large-scale learning to recover dense hand meshes with superior generalization. The former prioritizes model structure and estimation robustness, while the latter advances end-to-end learning performance and dataset scalability. Dong et al. [28] introduced a deep learning framework that bridges graph neural networks and state-space modeling for improved 3D hand pose and shape estimation from single red–green–blue (RGB) images. It addresses limitations in existing transformer-based models by proposing a Graph-guided State Space (GSS) block, which more efficiently captures spatial relationships between joints using significantly fewer tokens. Combined with a global-local fusion module, the proposed model achieves state-of-the-art performance on public benchmarks. In relation to this current study which offers a structured, estimation-driven method rooted in kinematic modeling and physical constraints, the study by Dong et al. [28] delivers a data-intensive, learning-driven solution that leverages graph-guided neural representations for superior accuracy. The former prioritizes robustness and model transparency, whereas the latter emphasizes scalability and predictive performance through architectural innovation.

Rezaei et al. [29] presented a deep learning-based method tailored for depth image inputs. Its key innovation lies in decomposing 3D pose estimation into two stages: predicting 2D joint locations (UV space) and separately estimating depth using dual attention maps. This decomposition isolates the complexity of depth prediction, improving estimation accuracy. Additionally, it introduces a novel appearance-based data augmentation technique for depth images. This current study emphasizes explicit kinematic modeling, interpretability, and estimation robustness without training, while the study by Rezaei et al. [29] focuses on architectural innovations and data augmentation to enhance learning-based prediction accuracy. The former is designed for model-driven reconstruction under minimal data assumptions using measured red–green–blue–depth (RGB-D) data, while the latter advances data-driven depth image analysis for applications.

Despite these advancements, current models are designed for robotic applications and are not readily adaptable to map real hand landmark point clouds into kinematic representations for accurate remote control and graphical reconstruction. In this study, hand poses were estimated using a kinematic model that explicitly enforces motion constraints and ensures natural hand configurations. The proposed approach presents a 3D objective function using a Bayesian tracking framework for enhancing robustness to occlusion and missing data by providing a well-constrained solution space that accurately reflects hand biomechanics. Additionally, deep learning-based hand pose reconstruction methods typically rely on implicit learning of hand structure constraints within the network, which makes them sensitive to unseen hand poses and limits their capacity to generalize across the full range of hand motion. In this study, a hierarchical kinematic model tailored for hand tracking applications was proposed. The proposed model integrates measured data with a multi-layered kinematic structure, enabling precise hand tracking and reconstruction that aligns well with real hand movements in 3D space.

2 Landmark-Based Hand Kinematic Construction

This section presents a short overview of kinematic modeling and construction of the hand based on tracking landmarks.

Kinematic modeling involves creating a mathematical framework to represent the motion of a physical structure, such as the human hand. This representation can be used to determine the relative joint variables that describe the position and orientation of various interconnected segments. Such information can further be utilized to control a physical robotic hand or its virtual counterpart, or as part of a graphical model for visualizing the physical system.

In contrast to physical robots or wearable devices - where each moving joint is equipped with sensors for precise motion tracking - hand movement tracking using ambient sensing modalities such as RGB-D sensors provides only the coordinates of selected landmarks relative to the sensor frame. For example, Figure 1 shows typical hand landmarks defined by MediaPipe, measured using calibrated RGB-D sensing (e.g., an Intel RealSense RGB-D sensor, D345-SDK), with coordinates defined in the sensor's coordinate system.

The goal of constructing a kinematic model of the hand is to establish a hierarchical coordinate system that can be used to compute finger joint angles based on these spatial landmark measurements through the inverse kinematic solutions. On the other hand, the forward solution refers to the case where the information regarding finger joint angles is available and it is required to estimate for the spatial location of landmarks. In this study, the kinematic model of the hand was constructed from 21 spatial landmark measurements (a point cloud with respect to the sensor frame), consisting of one at the root (the wrist) and 20 distributed across the phalanges of the five fingers. In the following kinematic model construction, each link was assigned a corresponding local frame, allowing the description of the position and rotational angles of each link.

A world coordinate frame $\{\mathrm{W}\}$ is defined to locate the position and orientation of the hand, including its palm and connecting fingers. This coordinate frame is collocated with the RGB-D sensor frame. The rotational angle of each finger joint can be described by defining local (or relative) coordinate frames within the hand. In this context, a wrist frame $\{0\}$ is defined, which is affixed to the wrist, moves with the hand, and is specified relative to the world frame. It also serves as a local coordinate frame with respect to which all other hand coordinates are defined. As shown in Figure 1, the $\hat{Y}_0$ of the wrist frame is defined to be aligned with the general direction of finger flexion, the unit vector $\hat{x}_0$ is oriented from the wrist towards the root of the middle finger, and $\hat{Z}_0$ is perpendicular to the palm plane, pointing towards the back of the hand, assuming that the front of the hand faces the sensor.

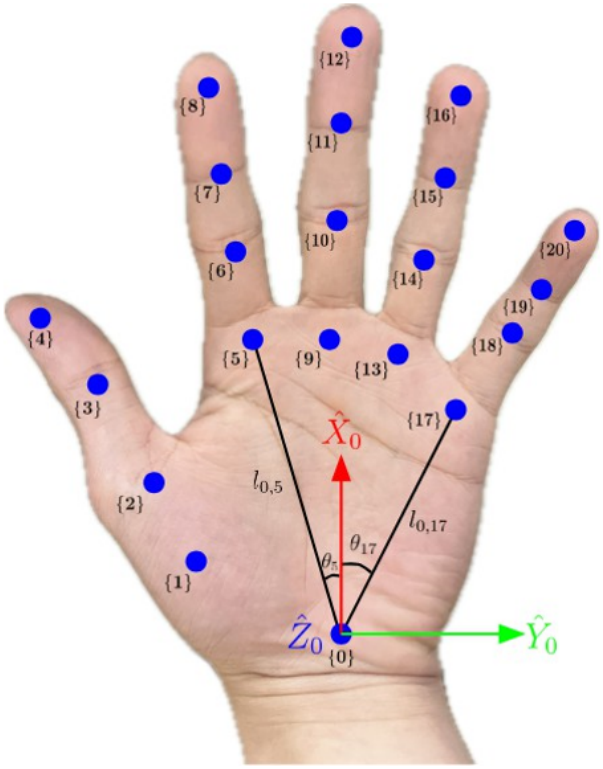

Figure 1. Hierarchical local frames $\{0\}$ to $\{20\}$ and parameters of the hand palm

The proposed model employs a hierarchical structure of local frames corresponding to anatomical joints. As shown in Figure 1, the first layer includes local frames $\{1\}, \{5\}, \{9\}, \{13\}$ and $\{17\}$ at the MCP joints, defined with respect to the root frame $\{0\}$. The second layer consists of local frames $\{2\}, \{6\}, \{10\}$, $\{14\}$ and $\{18\}$ at the proximal interphalangeal (PIP) joints, defined relative to the first layer. Similarly, frames $\{3\}, \{7\}, \{11\}, \{15\}$ and $\{19\}$ at the distal interphalangeal (DIP) joints form the third layer. Finally, the fourth layer comprises frames $\{4\}, \{8\}, \{12\}, \{16\}$ and $\{20\}$ at the fingertips.

Each MCP joint, or finger root joint, possesses two DOF, enabling it to flex and extend (bend up and down) as well as abduct and adduct (move side to side). When the root of a finger is fixed, and the finger is bent up and down, the PIP and DIP joints each exhibit one DOF. Consequently, the workspace (the area the fingertip can reach) of a finger is a 2D plane (Figure 2a). Each of the five fingers can thus be modeled as a kinematic chain comprising three links connected by three revolute joints, with parallel rotation axes. The unit vectors $\hat{Y}$ of the local frames of each finger are aligned with the axes of the revolute joints, as shown in Figure 2b. The unit vectors $\hat{\mathrm{X}}$ for each of the three joints orient towards the origin of the subsequent frame. For example, along the index finger, $\hat{\mathrm{X}}_5$ points to the origin of local frame $\{6\}$, $\hat{\mathrm{X}}_6$ to the origin of local frame $\{7\}$, and $\hat{\mathrm{X}}_7$ to the origin of local frame $\{8\}$. Note that since $\{8\}$ is located at the fingertip and has no DOF, and hence $\hat{\mathrm{X}}_8$ aligns with X_7. The unit vectors $\hat{\mathrm{Z}}$ of the local frames are assigned using the right-hand rule.

Palm parameters: In this study, the kinematic model of the palm is represented as a planar surface in 3D space. It is assumed that the wrist and the roots of the five finger landmarks lie on a single plane, specifically the xy-plane of the local frame $\{0\}$. A set of sixteen parameters are defined, where the initial six parameters are time-varying while the five parameters defined as lengths and the other five angle parameters are assumed to be constant:

- Position parameters: The position vector $\mathbf{d}_0^w = (x_0, y_0, z_0)^T$ specifies translational distance between the origins of the wrist frame $\{0\}$ and the world frame $\{\mathrm{W}\}$ along $\hat{\mathrm{X}}_\mathrm{w}$, $\hat{\mathrm{Y}}_\mathrm{w}$ and $\hat{\mathrm{Z}}_\mathrm{w}$, respectively.

(a) (b)

Figure 2. An example of the kinematic coordinate frame assignments: (a) frontal view of the extended index finger; (b) lateral view of the index finger

- Orientation parameters: The orientation of the local frame $\{0\}$ is defined by the Euler angles $\{\gamma_0, \beta_0, \alpha_0\}$, and these angles describe the sequential rotations around axes $\hat{Z}_0, \hat{Y}_0$ and $\hat{X}_0$.

- Length parameters: The lengths $\{l_{0,1}, l_{0,5}, l_{0,9}, l_{0,13}, l_{0,17}\}$ represent the fixed distances between the origin of the local frame $\{0\}$ and the origins of the frames $\{1\}, \{5\}, \{9\}, \{13\}$, and $\{17\}$. These measurements, in centimeters, are spatial relationships between the wrist and finger roots.

- Angle parameters: The angles $\{\theta_1, \theta_5, \theta_9, \theta_{13}, \theta_{17}\}$ describe the rotational offset between the $\hat{\mathrm{X}}_0$ axis and the axes $\hat{\mathrm{X}}_1, \hat{\mathrm{X}}_5, \hat{\mathrm{X}}_9, \hat{\mathrm{X}}_{13}$, and $\hat{\mathrm{X}}_{17}$, respectively. These angles, measured in radians, are pivotal for capturing the hand's natural articulation around the $\hat{\mathrm{Z}}_0$ axis.

Parameters of fingers: As shown in Figure 2, each finger requires three length parameters (as constraints) and four angle parameters (between 0 and 90 degrees) for defining its pose, amounting to a total of 35 parameters for all five fingers:

- Length parameters: A set of 15 parameters determines the fixed lengths of the three links $l_{i,i+1}, l_{i+1,i+2}$ and $l_{i+2,i+3}$ within each finger, where $i = 1, 5, 9, 13, 17$:

$\{(l_{1,2}, l_{2,3}, l_{3,4}), (l_{5,6}, l_{6,7}, l_{7,8}), (l_{9,10}, l_{10,11}, l_{11,12}), (l_{13,14}, l_{14,15}, l_{15,16}), (l_{17,18}, l_{18,19}, l_{19,20})\}$

- Angle parameters for finger roots: A set of 10 parameters describes five pairs of rotation angles β_i and ψ_i around axes $\hat{\mathrm{Y}}_i$ and $\hat{\mathrm{Z}}_i$ for the root joints of each finger, where $i = 1, 5, 9, 13, 17$:

$\{(\beta_1, \gamma_1), (\beta_5, \gamma_5), (\beta_9, \gamma_9), (\beta_{13}, \gamma_{13}), (\beta_{17}, \gamma_{17})\}$

- Angle parameters for other joints: A set of 10 parameters describes five pairs of rotation angles β_{i+1} and β_{i+2} for the PIP and DIP joints around corresponding $\hat{\mathrm{Y}}_i$ of each finger, where $i = 1, 5, 9, 13, 17$:

$\{(\beta_2, \beta_3), (\beta_6, \beta_7), (\beta_{10}, \beta_{11}), (\beta_{14}, \beta_{15}), (\beta_{19}, \beta_{20})\}$

To effectively track human hand movements and reconstruct corresponding poses in a virtual environment, it is essential to represent the position and orientation of each joint's local coordinate frame in 3D space, along with the corresponding joint parameters of each finger. A systematic approach was introduced for resolving these parameters through transformations between adjacent coordinate frames. The input parameters are 21 hand landmarks provided by the MediaPipe model, along with depth data captured by an RGB-D sensor. These represent the position parameters of the origin of each local frame {i} with respect to world frame {W}, denoted as the vector $\mathbf{d}_i^w = (x_i, y_i, z_i)^T$ for $i = 0, 1, \ldots, 20$. Table 1 shows the definitions of these measured 3D spatial coordinates for each joint of the thumb, index, middle, ring, and little fingers.

Table 1. Definitions of the 3D spatial coordinates $\mathbf{d}_i^w$ of 21 hand landmarks (calibrated measures in meters)

	Joints			
Finger	**MCP (1st layer)**	**PIP (2nd layer)**	**DIP (3rd layer)**	**Tip (4th layer)**
Wrist		$\mathbf{d}_0^w$		
Thumb	$\mathbf{d}_1^w$	$\mathbf{d}_2^w$	d_3^w	$\mathbf{d}_4^w$
Index	$\mathbf{d}_5^w$	$\mathbf{d}_6^w$	$\mathbf{d}_7^w$	$\mathbf{d}_8^w$
Middle	$\mathbf{d}_9^w$	$\mathbf{d}_{10}^w$	$\mathbf{d}_{11}^w$	$\mathbf{d}_{12}^w$
Ring	$\mathbf{d}_{13}^w$	$\mathbf{d}_{14}^w$	$\mathbf{d}_{15}^w$	$\mathbf{d}_{16}^w$
Little	$\mathbf{d}_{17}^w$	$\mathbf{d}_{18}^w$	$\mathbf{d}_{19}^w$	$\mathbf{d}_{20}^w$

Table 2 provides an example of the 3D hand landmark data $\mathbf{d}_i^w$ in meters relative to the world frame, with its

axes aligned with the axes of the RGB-D sensor frame, (i.e., the positive y-axis direction pointing downward and the positive z-direction measuring the distance of the hand to the sensor).

Table 2. An example of calibrated measures of spatial coordinates $\mathbf{d}_i^w$ of 21 hand landmarks

	Joints			
Finger	**MCP (1st layer)**	**PIP (2nd layer)**	**DIP (3rd layer)**	**Tip (4th layer)**
Wrist	(-0.21, -0.11, 0.54)			
Thumb	(-0.19, -0.15, 0.53)	(-0.17, -0.18, 0.54)	(-0.15, -0.20, 0.55)	(-0.14, -0.22, 0.56)
Index	(-0.13, -0.16, 0.56)	(-0.10, -0.16, 0.56)	(-0.08, -0.16, 0.55)	(-0.06, -0.16, 0.55)
Middle	(-0.13, -0.14, 0.57)	(-0.09, -0.14, 0.57)	(-0.07, -0.14, 0.56)	(-0.05, -0.14, 0.55)
Ring	(-0.13, -0.12, 0.57)	(-0.10, -0.12, 0.57)	(-0.07, -0.12, 0.56)	(-0.06, -0.12, 0.56)
Little	(-0.14, -0.10, 0.57)	(-0.11, -0.10, 0.57)	(-0.09, -0.10, 0.57)	(-0.07, -0.10, 0.57)

Figure 3 shows the 2D hand tracking data visualization of the hand landmarks and the plot of calibrated 3D coordinates of the landmarks. In this figure, the RGB image shows a hand in a front-facing position, where key hand landmarks are tracked using the MediaPipe framework. These landmarks are accurately projected onto the image plane, ensuring comprehensive coverage of all joint locations. On the right, the 3D point cloud of these landmarks is visualized using depth data from an RGB-D sensor (Intel RealSense D345). The depth values associated with each landmark are mapped to their respective x and y coordinates derived from the MediaPipe output.

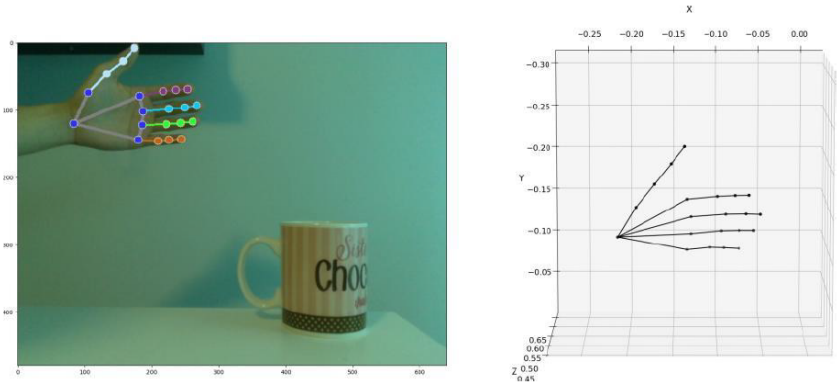

Figure 3. 2D pixels of MediaPipe hand tracking and 3D spatial point visualization

3 Estimation with Gaussian Bayesian Network

The stages of grasping, as captured using RGB-D and MediaPipe, are illustrated in Figure 4. This figure provides example case studies showing how the hand approaches, contacts, partially encloses, and ultimately fully grasps the object. These stages are depicted from two different measurement perspectives: a frontal view (top row) and a side view (bottom row). It is important to note that during all these phases, the positions and orientations of certain landmarks are more difficult to track and kinematically reconstruct accurately due to several factors. Occlusions - caused by parts of the hand being hidden behind the object or by overlapping fingers - can result in undetectable measures in depth values. Additionally, sensor data may be noisy or incomplete, particularly for joints involved in fine-grained grasping [30], such as the PIP and DIP joints. Missing or noisy measurements of these joints pose significant challenges for accurately reconstructing the full hand posture, as even small inaccuracies can lead to incorrect reconstruction of the grasp.

In the context of kinematic hand modeling for grasping objects, accurately estimating parameters such as joint angles for graphical reconstruction requires capturing the complex dependencies between adjacent joints. These natural dependencies are due to the structure of the open-kinematic chain representing each finger which is attached to the palm. In this construction, the movement of one joint influences the movement of others in the process of forming a stable grasp. To effectively model these relationships and account for uncertainties in sensor data, a Bayesian network framework was adopted. A Bayesian network is a probabilistic graphical model (PGM) that represents a set of variables and their conditional dependencies through a directed acyclic graph (DAG). In this framework, each node corresponds to a joint or landmark, and the edges capture the dependencies between joints, making it particularly well-suited for tracking a grasping hand where resolving the values of certain joints is more prone to occlusion.

This section extends this concept to a Gaussian Bayesian Network (GBN) [31], which is also well-suited for continuous domains such as joint positions and angles. The GBN framework allows us to model the uncertainty in

sensor data through Gaussian noise while capturing the hierarchical dependencies between joints in the kinematic structure and solutions of the hand. In the GBN framework, each joint position is represented as a continuous random variable, and the joint probability density function is modeled using a Gaussian distribution. The core strength of the Bayesian network lies in its ability to model these conditional dependencies between joints. For instance, given the known position of the wrist, the positions of the MCP joints are conditionally dependent on the position and orientation of the wrist. Similarly, the PIP and DIP joints depend on the state of their parent joints. This hierarchical relationship allows the network to propagate information efficiently through the kinematic model and solutions, making it possible to estimate joint angles even when some data points are missing or noisy.

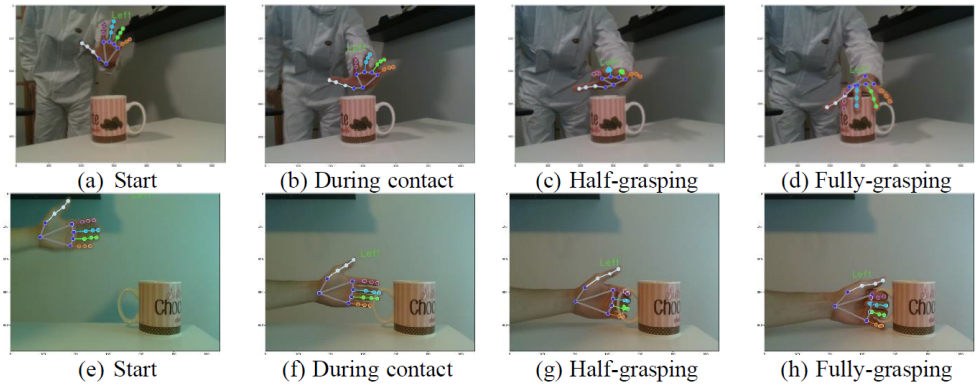

Figure 4. Stages of grasping measured from two perspectives using a hand-tracking system

Figure 4 shows the frontal and side measurement stages, respectively. Each frame overlays a detailed hand skeleton, depicting joint landmarks and tracked kinematic motion to demonstrate dynamic hand-object interaction.

3.1 Parameters of Gaussian Bayesian Network

As depicted in Figure 5, the root node in the network corresponds to the 3D spatial position of the origin of the wrist coordinate frame $\{0\}$ with respect to the world frame $\{\mathrm{W}\}$. This is obtained by the measured distance $\mathbf{d}_0^w$ of landmark 0, obtained from the RGB-D sensor and through calibrated measure of the MediaPipe's landmark and its corresponding depth measures. Each subsequent node in the network corresponds to the position of origin of a specific hand joint coordinate frame $\{\mathrm{j}\}, j = 1, \ldots, 20$, with respect to its parent coordinate frame $\{\mathrm{i}\}$, measured by the corresponding landmark j, with edges capturing the conditional dependencies between coordinate frames of the joints. The network structure reflects the hierarchical coordinate frame definitions for the hand kinematic model. For example, each position of the origin of child joint coordinate frames (e.g., PIP joints, $\{2\}, \{6\}, \{10\}, \{14\}, \{18\}$ in the second layer), is conditioned on the position of origin of its parent joint coordinate frames (e.g., MCP joints, $\{1\}$, $\{5\}, \{9\}, \{13\}, \{17\}$ at the first layer). This is the dependency structure so that each joint's spatial relationship is computed relative to its parent in the kinematic chain.

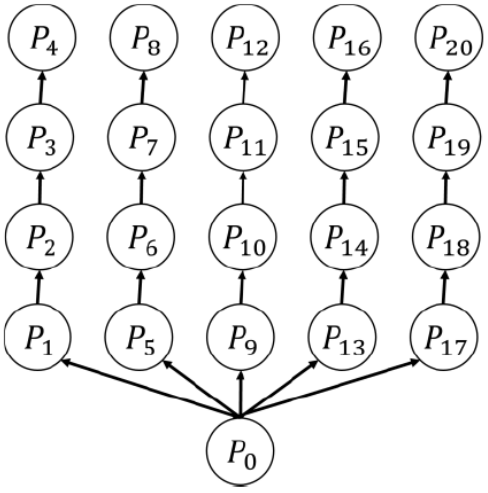

Figure 5. Bayesian network for 21 hand landmarks

When measured landmark information is available, a forward kinematic solution (or direct kinematics) is applied to estimate and perform kinematic transformations for reconstructing hand joint positions and orientations based on the measured landmark parameters. In cases where detected landmarks are missing or occluded by objects or other parts of the hand, forward estimation alone may yield inaccurate joint angle parameters, resulting in failed graphical reconstructions in Unity. To address such scenarios, an inverse kinematic solution was applied to resolve and identify an optimal set of hand kinematic model parameters. This method corrects and updates the joint angles by refining the forward kinematic model to better fit the available, incomplete data. By leveraging the forward kinematic solution-based on detected hand joint positions and previously estimated joint angles-iterative estimation enables the accurate reconstruction of hand movements and object grasping poses under various conditions. As illustrated in Figure 6, significant outliers typically have unusually large deviations from the other landmarks. A straightforward yet effective method for outlier detection is to compute the median of the 3D coordinates across all 21 landmarks and then measure the distance of each landmark from this median. If the distance exceeds a predefined threshold (e.g., 0.2 meters), the landmark is classified as an outlier.

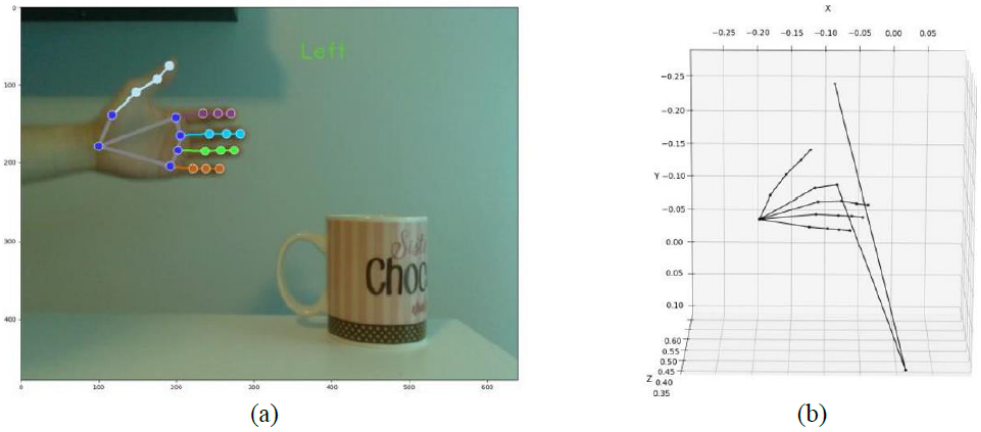

Figure 6. Visulization of hand tracking landmarks and its kinematic model reconstruction: (a) MediaPipe 2D landmark measures; (b) 3D kinematic model with the presence of the third joint depth outlier

The noise model of the sensed measured information of Intel RealSense D435 has been verified to follow a Gaussian (normal) distribution $N\left(0,\sigma^2\right)$ [32, 33] to account for the uncertainty in the depth sensor's measurements. This is for a given measured position vector $\mathbf{d}_j^w = (x_j, y_j, z_j)$ of landmark j, obtained from MediaPipe and the depth sensor. This measure can be interpreted to follow a Gaussian distribution, with the mean representing the "kinematic" position of the landmark (the expected or theoretical position of a landmark based on the kinematic chain model of the hand) and the variance:

$$\mathbf{d}_j^w \sim \mathrm{N}\left(\mathrm{T}_j^i\left(\mathbf{d}_i^w, \Theta_{i,j}, l_{i,j}\right), \sigma^2\right), \tag{1}$$

where, $\mathrm{T}_j^i\left(\mathbf{d}_i^w, \Theta_{i,j}, l_{i,j}\right)$ is the expected 3D "kinematic" position of the landmark j, computed based on the hierarchical kinematic transformations (defined in the previous section) from its parent joint $\mathbf{d}_i^w$, the joint angle parameter(s) Θ_{ij} (e.g., $\Theta_{w,0} = (\alpha_0, \beta_0, \gamma_0) \in [0, 2\pi]$, for the first layer joints $i = 1, 5, 9, 13, 17, \Theta_{0,i} = \theta_i$ and $\Theta_{i,i+1} = (\beta_i, \gamma_i) \in \left[0, \frac{\pi}{2}\right]$ and for the second layer joints $j = 2, 6, 10, 14, 18, \Theta_{j,j+1} = \beta_j \in \left[0, \frac{\pi}{2}\right]$ and $\Theta_{j+1,j+2} = \beta_{j+1} \in \left[0, \frac{\pi}{2}\right]$), and the fixed link (finger) length $l_{i,j,j}$, which serves as a constraint to define the range for the estimation process of angle parameters in the next section.

These parameters in $\mathrm{T}_j^i\left(\mathbf{d}_i^w, \Theta_{i,j}, l_{i,j}\right)$ define the hand's spatial configuration, including the orientation and physical length of each segment between joints. Through a series of hierarchical transformations, which are based on the principles of forward kinematics, the expected 3D spatial position of landmark j can be calculated.

In contrast, the actual measured position of the landmark from the sensor may deviate from this expected "kinematic" position due to various factors, such as the uncertainty and noise in the depth sensor's measurements or occlusions. By modeling the observed position as a Gaussian distribution centered around the expected "kinematic" position, this uncertainty is accounted for, and the joint angles can be estimated more accurately by comparing measured positions to their corresponding kinematic positions. The probability of observing a 3D position $\mathbf{d}_j^w$, conditioned on the hierarchical kinematic transformation based on its parent landmark $\mathbf{d}_i^w$, the joint angle parameter(s) $\Theta_{i,j}$, and the length l_{ij}, is then represented by the following probability density function:

$$P\left(\mathbf{d}_j^w \mid \mathrm{T}_j^i\left(\mathbf{d}_i^w, \Theta_{i,j}, l_{i,j}\right), \sigma\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{\left\|\mathbf{d}_j^w - \mathrm{T}_j^i\left(\mathbf{d}_i^w, \Theta_{i,j}, l_{i,j}\right)\right\|^2}{2\sigma^2}\right\} \tag{2}$$

Ideally, in the absence of noise, each measured position $\mathbf{d}_j^w$ would be exactly equal to the kinematic position given by the transformation $\mathrm{T}_j^i\left(\mathbf{d}_i^w, \Theta_{i,j}, l_{ij}\right)$, leading to a probability of 1. However, sensor noise is inevitable, meaning the measured data will always deviate from the ideal kinematic positions. For instance, as illustrated in Figure 6, the depth values for the third and the fingertip landmarks of the index finger (joints 7 and 8) are particularly inaccurate. Therefore, the distance computed distance $\left\|\mathbf{d}_7^w - \mathrm{T}_7^6\left(\mathbf{d}_6^w, \beta_6, l_{6,7}\right)\right\|$ where $\mathrm{T}_7{}^6\left(\mathbf{d}_6^w, l_{6,7}, \Theta_{6,7}\right)$ is the estimated position of landmark 7 derived from the kinematic model using the measured position of its parent landmark $\mathbf{d}_6^w$, the resolved finger joint angle $\Theta_{6,7} = \beta_6$, and length $l_{6,7}$ as a constraint - remains large due to noise in the measured position $\mathbf{d}_7^w$. This leads to the following exponential term approaching zero:

$$\exp\left\{-\frac{\left\|\mathbf{d}_7^w - \mathrm{T}_7^6\left(\mathbf{d}_6^w, \beta_6, l_{6,7}\right)\right\|_2^2}{2\sigma^2}\right\}. \tag{3}$$

As a result, probability $P\left(\mathbf{d}_7^w \mid \mathrm{T}_7^6\left(\mathbf{d}_6^w, \beta_6, l_{6,7}\right), \sigma\right)$ is close to 0, indicating a poor match between the kinematic model and measured positions. A similar situation occurs for joint 8, where $P\left(\mathbf{d}_8^{\mathrm{w}} \mid \mathrm{T}_8^7\left(\mathbf{d}_7^{\mathrm{w}}, \beta_7, l_{7,8}\right), \sigma\right)$. In the presence of such anomalies described above, this study presents an optimal estimation of kinematic model parameters which enables, in particular, the resolution of the joint angle parameters $\Theta_{i,j}$ that best explain (or fit) the measured data corresponding to hand movements.

3.2 Method of Maximum Likelihood Estimation

To estimate the optimal set of parameters $\Theta_{i,j}$ and $\mathbf{d}_i^w$ that explain the observed data, Maximum Likelihood Estimation (MLE) was employed [34]. The likelihood function is the joint probability of observing all the measured landmark positions, assuming the parameters $\Theta_{i,j}$ and the fixed link lengths $l_{i,j}$ are known. Given the measurement noise model, the likelihood function for a single joint observation $\mathbf{d}_j^w$ can be expressed as:

$$L\left(\mathrm{T}_j^i\left(\mathbf{d}_i^w, \Theta_{i,j}, l_{i,j}\right), \sigma \mid \mathbf{d}_j^w\right) = P\left(\mathbf{d}_j^w \mid \mathrm{T}_j^i\left(\mathbf{d}_i^w, \Theta_{i,j}, l_{i,j}\right), \sigma\right). \tag{4}$$

The goal of MLE is to find the parameter values that maximize the likelihood of the observed data, which, in this case, are the measured 3D positions of the hand landmarks $\mathbf{d}_j^w$ obtained from the sensor. Since there are 21 landmark measurements (including landmark 0) defined by MediaPipe and obtained through the Intel D435 RGB-D depth sensor, this study aims to estimate how well the optimal sets of parameters for the hand kinematic model describe the hand pose during movement. To accomplish this, this study seeks to maximize the overall likelihood, which is the product of the individual probabilities for each observation. Mathematically, this can be represented as the joint likelihood of all 21 landmarks in each frame:

$$\begin{aligned}
&L\left(\mathrm{T}(\mathbf{d}, \Theta, l), \sigma \mid \mathbf{d}_0^w, \mathbf{d}_1^w, \cdots, \mathbf{d}_{20}^w\right) = P\left(\mathbf{d}_0^w, \mathbf{d}_1^w, \cdots, \mathbf{d}_{20}^w \mid \mathrm{T}(\mathbf{d}, \Theta, l), \sigma\right) \\
&= P\left(\mathbf{d}_0^w \mid \mathrm{T}\left(\mathbf{d}_w^w, (\alpha_0, \beta_0, \gamma_0), l_{w,0}\right), \sigma\right) \cdot P\left(\mathbf{d}_1^w \mid \mathrm{T}\left(\mathbf{d}_0^w, \theta_1, l_{0,1}\right), \sigma\right) \cdot P\left(\mathbf{d}_5^w \mid \mathrm{T}\left(\mathbf{d}_0^w, \theta_5, l_{0,5}\right), \sigma\right) \\
&\cdot P\left(\mathbf{d}_9^w \mid \mathrm{T}\left(\mathbf{d}_0^w, \theta_9, l_{0,9}\right), \sigma\right) \cdot P\left(\mathbf{d}_{13}^w \mid \mathrm{T}\left(\mathbf{d}_0^w, \theta_{13}, l_{0,13}\right), \sigma\right) \cdot P\left(\mathbf{d}_{17}^w \mid \mathrm{T}\left(\mathbf{d}_0^w, \theta_{17}, l_{0,17}\right), \sigma\right) \\
&\cdot P\left(\mathbf{d}_2^w \mid \mathrm{T}\left(\mathbf{d}_1^w, (\beta_1, \gamma_1), l_{1,2}\right), \sigma\right) \cdot P\left(\mathbf{d}_3^w \mid \mathrm{T}\left(\mathbf{d}_2^w, \beta_2, l_{2,3}\right), \sigma\right) \cdot P\left(\mathbf{d}_4^w \mid \mathrm{T}\left(\mathbf{d}_3^w, \beta_3, l_{3,4}\right), \sigma\right) \\
&\cdot P\left(\mathbf{d}_6^w \mid \mathrm{T}\left(\mathbf{d}_5^w, (\beta_5, \gamma_5), l_{5,6}\right), \sigma\right) \cdot P\left(\mathbf{d}_7^w \mid \mathrm{T}\left(\mathbf{d}_6^w, \beta_6, l_{6,7}\right), \sigma\right) \cdot P\left(\mathbf{d}_8^w \mid \mathrm{T}\left(\mathbf{d}_7^w, \beta_7, l_{7,8}\right), \sigma\right) \\
&\cdot P\left(\mathbf{d}_{10}^w \mid \mathrm{T}\left(\mathbf{d}_9^w, (\beta_9, \gamma_9), l_{9,10}\right), \sigma\right) \cdot P\left(\mathbf{d}_{11}^w \mid \mathrm{T}\left(\mathbf{d}_{10}^w, \beta_{10}, l_{10,11}\right), \sigma\right) \cdot P\left(\mathbf{d}_{12}^w \mid \mathrm{T}\left(\mathbf{d}_{11}^w, \beta_{11}, l_{11,12}\right), \sigma\right) \\
&\cdot P\left(\mathbf{d}_{14}^w \mid \mathrm{T}\left(\mathbf{d}_{13}^w, (\beta_{13}, \gamma_{13}), l_{13,14}\right), \sigma\right) \cdot P\left(\mathbf{d}_{15}^w \mid \mathrm{T}\left(\mathbf{d}_{14}^w, \beta_{14}, l_{14,15}\right), \sigma\right) \cdot P\left(\mathbf{d}_{16}^w \mid \mathrm{T}\left(\mathbf{d}_{15}^w, \beta_{15}, l_{15,16}\right), \sigma\right) \\
&\cdot P\left(\mathbf{d}_{18}^w \mid \mathrm{T}\left(\mathbf{d}_{17}^w, (\beta_{17}, \gamma_{17}), l_{17,18}\right), \sigma\right) \cdot P\left(\mathbf{d}_{19}^w \mid \mathrm{T}\left(\mathbf{d}_{18}^w, \beta_{18}, l_{18,19}\right), \sigma\right) \cdot P\left(\mathbf{d}_{20}^w \mid \mathrm{T}\left(\mathbf{d}_{19}^w, \beta_{19}, l_{19,20}\right), \sigma\right) \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) \exp\left\{-\frac{\left\|\mathbf{d}_0^w - \mathrm{T}_0^w\left(\mathbf{0}, \Theta_{w,0}, l_{w,0}\right)\right\|^2}{2\sigma^2}\right\} \cdot \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) \exp\left\{-\frac{\left\|\mathbf{d}_1^w - \mathrm{T}_1^0\left(\mathbf{d}_0^w, \Theta_{0,1}, l_{0,1}\right)\right\|^2}{2\sigma^2}\right\} \\
&\cdot \cdots \cdot \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) \exp\left\{-\frac{\left\|\mathbf{d}_{20}^w - \mathrm{T}_{20}^{19}\left(\mathbf{d}_{19}^w, \Theta_{19,20}, l_{19,20}\right)\right\|^2}{2\sigma^2}\right\} \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{21} \exp\left\{-\frac{1}{2\sigma^2} \sum_{j=0}^{20} \left\|\mathbf{d}_j^w - \mathrm{T}_j^i\left(\mathbf{d}_i^w, \Theta_{i,j}, l_{i,j}\right)\right\|^2\right\}.
\end{aligned} \tag{5}$$

In practice, the log-likelihood function, which converts the product of probabilities into a sum of log-probabilities, is maximized to simplify computation. This transformation allows us to more easily manage the complexity of the estimation process:

$$\begin{aligned}
&\ln\left[L\left(\mathrm{T}(\mathbf{d},\Theta,l),\sigma \mid \mathbf{d}_0^w,\mathbf{d}_1^w,\cdots,\mathbf{d}_{20}^w\right)\right] = \sum_{j=0}^{20} \ln\left\|P\left(\mathbf{d}_j^w \mid \mathrm{T}_j^i\left(\mathbf{d}_i^w,\Theta_{i,j},l_{i,j}\right),\sigma\right)\right\| \\
&= \ln\left(\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{21}\exp\left\{-\frac{1}{2\sigma^2}\sum_{j=0}^{20}\left\|\mathbf{d}_j^w-\mathrm{T}_j^i\left(\mathbf{d}_i^w,\Theta_{i,j},l_{i,j}\right)\right\|^2\right\}\right) \\
&= \ln\left(\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{21}\right)+\ln\left(\exp\left\{-\frac{1}{2\sigma^2}\sum_{j=0}^{20}\left\|\mathbf{d}_j^w-\mathrm{T}_j^i\left(\mathbf{d}_i^w,\Theta_{i,j},l_{i,j}\right)\right\|^2\right\}\right) \\
&= -\frac{21}{2}\ln\left(2\pi\sigma^2\right)-\frac{1}{2\sigma^2}\sum_{j=0}^{20}\left\|\mathbf{d}_j^w-\mathrm{T}_j^i\left(\mathbf{d}_i^w,\Theta_{i,j},l_{i,j}\right)\right\|^2.
\end{aligned} \tag{6}$$

This study further assumes that σ remains constant and maximizing the second term of the log-likelihood function can only be focused on:

$$-\frac{1}{2\sigma^2}\sum_{j=0}^{20}\left\|\mathbf{d}_j^w-\mathrm{T}_j^i\left(\mathbf{d}_i^w,\Theta_{i,j},l_{i,j}\right)\right\|^2 \tag{7}$$

This translates into minimizing the sum of squared errors (SSE) between the observed landmark positions $\mathbf{d}_j^w$ and the predicted kinematic positions $T_j^i\left(\mathbf{d}_i^w,\Theta_{i,j},l_{i,j}\right)$:

$$\sum_{j=0}^{20}\left\|\mathbf{d}_j^w-\mathrm{T}_j^i\left(\mathbf{d}_i^w,\Theta_{i,j},l_{i,j}\right)\right\|^2. \tag{8}$$

To find the optimal set of angle parameters Θ and position vectors d, the error in the predicted kinematic model is minimized by solving the least squares problem presented in Eq. (8). The objective is to reduce the discrepancy between the measured and model positions, ensuring the estimated parameters fit the observed hand movements as closely as possible. However, as it was mentioned before, in practical applications, hand landmark detection can be prone to failures or erroneous results due to factors such as occlusions, noise effects, or sensor limitations. These issues may produce outliers in depth measurements - such as zero values (in object-occluded or self-occluded cases) or incorrectly assigned background values (e.g., from a background white wall). Depth measurement inaccuracies directly impact the transformation from 2D pixel coordinates $p=(u_i,v_i)$ to 3D spatial coordinates $d_i=(x_i,y_i,z_i)$, as this conversion relies critically on accurate depth data $d\left(u_i,v_i\right)$ using intrinsic camera parameters, specifically focal lengths f_x and f_y and principal point offsets c_x and c_y:

$$\begin{aligned}
x_i &= \frac{\left(u_i-c_x\right)\cdot d\left(u_i,v_i\right)}{f_x} \\
y_i &= \frac{\left(v_i-c_y\right)\cdot d\left(u_i,v_i\right)}{f_y} \\
z_i &= d\left(u_i,v_i\right), \quad i=0,\ldots,20
\end{aligned} \tag{9}$$

Minimizing the 3D coordinate error in such cases can lead to suboptimal parameter estimates, especially when multiple outliers are present. To address these outliers, a binary confidence score $o_j\in\{0,1\}$ was introduced for each landmark j. This score indicates whether the measured landmark is an outlier, where $o_j=1$ denotes an outlier and $o_j=0$ represents a reliable measurement. By incorporating this confidence score into the proposed optimization process, unreliable landmarks can be selectively excluded from influencing the parameter estimation. Then the modified objective function to handle these outliers is then expressed as:

$$\sum_{j=0}^{20}\left(1-o_j\right)\left\|\mathbf{d}_j^w-\mathrm{T}_j^i\left(\mathbf{d}_i^w,\Theta_{i,j},l_{i,j}\right)\right\|^2. \tag{10}$$

4 Qualitative Experimental Study

This section presents sample qualitative results for two sets of experiments involving tracking and graphical reconstruction sequences of the hand movements toward a cup for grasping while the measured observations were conducted from two different perspectives. These experiments are designed to test the robustness of the proposed estimation approach under varying conditions, including various cases of occlusions. The tracking and reconstruction were initialized using the kinematic model parameters in order to calculate the initial joint angles and position of the virtual hand model (Figure 7). Initially, the parameters ($\mathbf{d}_i^w$ for $i = 0, \ldots, 20$) are captured by MediaPipe and the depth sensor, while the link lengths l_{ij} are fixed. The joint angle parameters Θ_{ij} can either be initialized to zero (if no prior information is available, e.g., the starting point of a hand movement) or set based on prior knowledge (e.g., the estimated hand pose from the previous frame). In the case studies of this research, hand movements were recorded as sequences of RGBdepth images to analyze different motions. As for the computation of coordinates, the current parameters ($\mathbf{d}_i^w, \Theta_{i,j}$, and $l_{i,j}$) were used for each parent landmark i, the hierarchical transformations $\mathrm{T}_j^i\,(\mathbf{d}_i^w, \Theta_{i,j}, l_{i,j})$ were performed to compute the estimated $\mathbf{d}_j^w$ for each child landmark of the hand. The accuracy of the current parameter estimation was evaluated by comparing the estimated coordinate $\mathbf{d}_j^w$ with the measured coordinates obtained from the MediaPipe hand model and the RGB-D sensor. Finally, the objective function's value (or "loss") was used to optimize the model parameters using a gradient-based method.

The first grasping task case study, in which the hand approaches a cup from above in a front-facing view, is shown in Figure 8 and Figure 9. This study presents the results of hand tracking and reconstruction during the initial approach phase, followed by two intermediate pre-grasping poses, and finally the full grasping phase. A second grasping task case study, shown in Figure 10 and Figure 11, involves the hand moving along a diagonal trajectory from the upper left to the lower right with respect to the sensor. In this case, tracking data is collected from a side view, culminating in a full grasp of the cup handle. Using a similar analysis to the previous experiment, the performance of the estimation method was evaluated along this alternative trajectory and viewpoint, with a focus on the model's adaptability to changes in perspective and motion path.

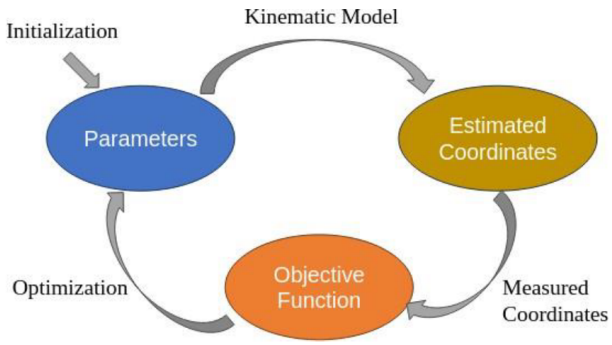

Figure 7. Schematic diagram of the model parameter optimization process

Figures 8a, c, e, g show RGB hand motion tracking as the hand approaches the cup: directly facing the sensor, during contact, partial grasp with self-occlusion, and full grasp with heavy occlusion. Figures 8b, d, f, h show the measured (black) and estimated (purple) 3D spatial landmark points.

Unity graphical reconstructions using forward kinematics are shown in Figures 9a, c, e, g, and refined Unity graphical reconstructions with estimation are shown in Figures 9b, d, f, h.

Figures 10a, c, e, g show RGB hand motion tracking as the hand approaches the cup: directly facing the sensor, during contact, partial grasp with self-occlusion, and full grasp with heavy occlusion. Figures 10b, d, f, h show the measured (black) and estimated (purple) 3D spatial landmark points.

Unity graphical reconstructions using forward kinematics are shown in Figures 11a, c, e, g, and refined Unity graphical reconstructions with estimation are shown in Figures 11b, d, f, h.

The following are observed in regard to the case studies presented in Figure 8 and Figure 9 (Case 1) and Figure 10 and Figure 11 (Case 2). For the case of a hand approaching and grasping a cup from the top while collecting tracking measurements using the RGB-D sensor from a frontal view of the hand, the following are observed for the selected sequences stated in Figure 8. In non-occlusion case 1 (α Instant), at the initial state of motion, three outliers were detected for the fingertip landmarks. In non-occlusion case 1 (β Instant), during the pregrasping phases, nine outliers were detected. In occlusion case 1 (γ Instant), self-occlusion occurred, resulting in missing ring finger landmark measurements. In occlusion case 1 (δ Instant), self-occlusion was detected for missing landmark readings of the palm. In the case of the hand approaching and grasping the cup by its handle, while RGB-D tracking measurements are obtained from the side view of the hand (Figure 10), the following instances are selected and analyzed. In

non-occlusion case 2 (α Instant) when the hand is at its initial motion configuration, three landmark outliers were detected. In object-occlusion case 2 (β Instant), when the fingertips are touching the cup handle, landmark outliers were detected. In occlusion case 2(γ Instant), when the fingers are grasping the handle, ambiguous fingertip landmark measures were observed. In occlusion case 2 (δ Instant), when the handle of the cup is fully grasped, ambiguous finger landmark measures were present. The following presents detailed analysis and discussions for two sample instances mentioned above (Figure 8 and Figure 10).

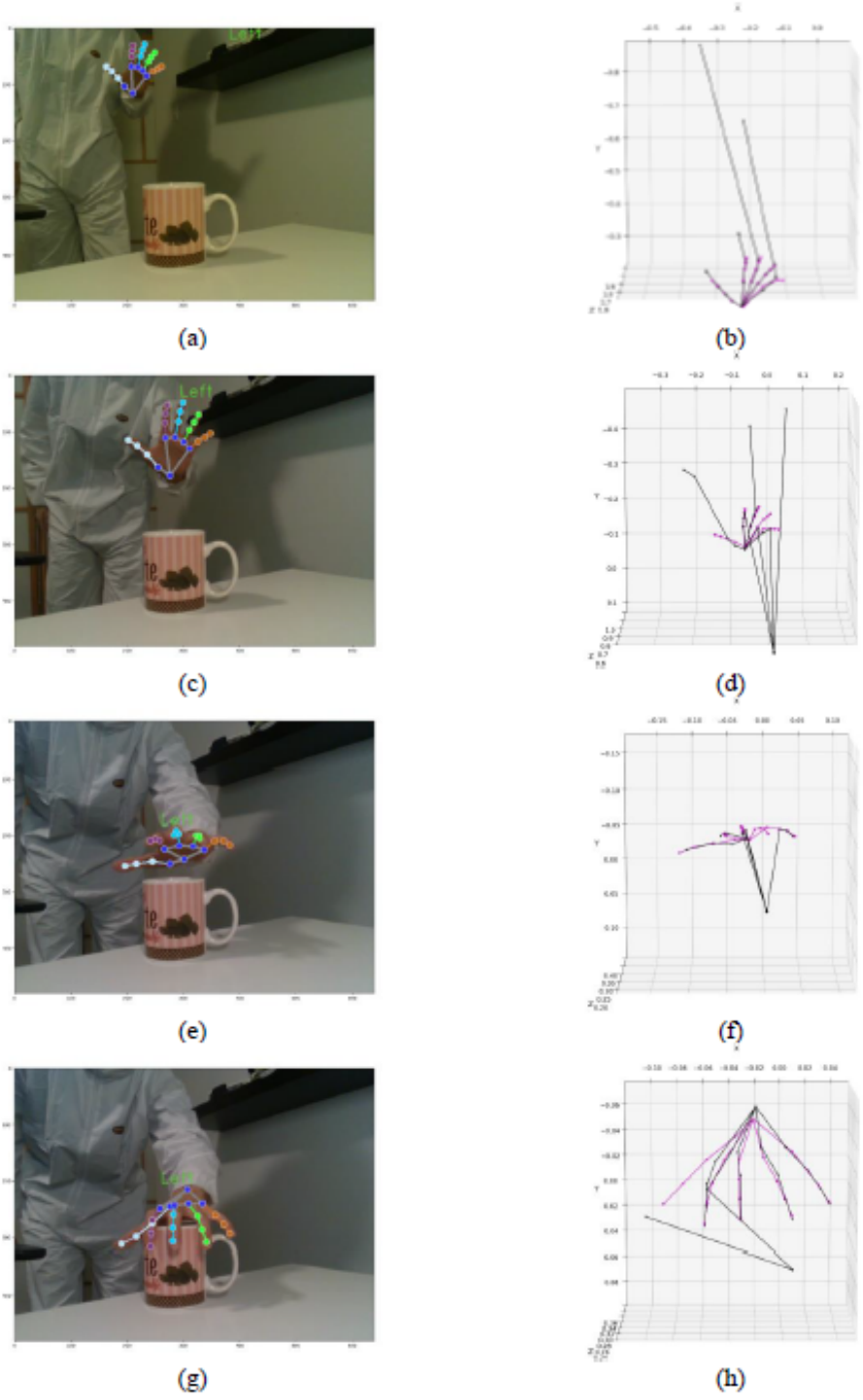

Figure 8. Hand tracking sequences in the frontal view: (a) α Instant – landmarks, (b) α Instant – measurements, (c) β Instant – landmarks, (d) β – measurements, (e) γ Instant – landmarks, (f) γ Instant – measurements, (g) δ Instant – landmarks, (h) δ Instant - measurements

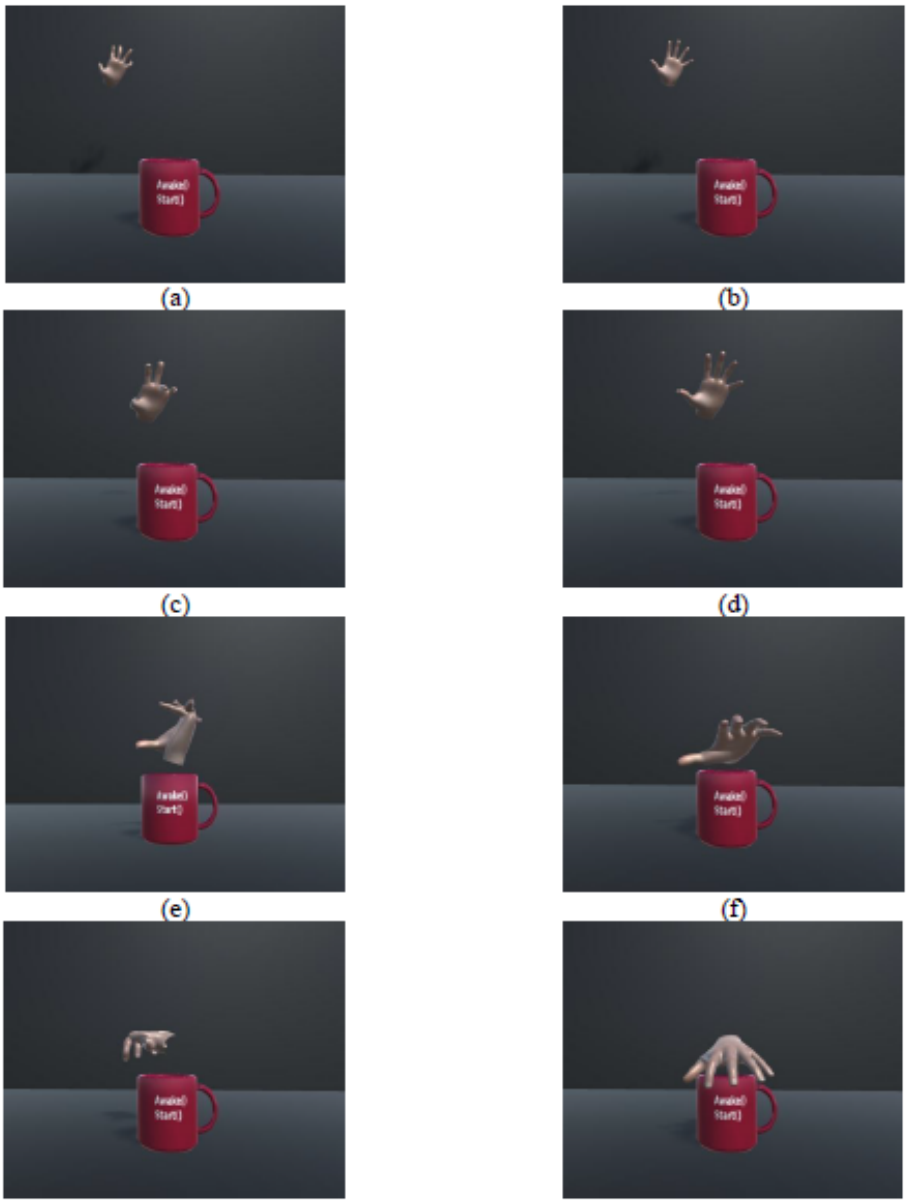

Figure 9. Unity hand graphical reconstruction associated with the sequences of Figure 8: (a) α Instant – original, (b) α Instant – refined, (c) β Instant – original, (d) β Instant – refined, (e) γ Instant – original, (f) γ Instant – refined, (g) δ Instant – original, (h) δ Instant - refined

4.1 γ Instant of the Frontal Measure: Case 1

This section presents some additional details associated with the γ instance of the hand tracking and reconstruction in the frontal measure case (Figure 8). This case, as stated above, involves self-occlusion of the hand with a missing measure of the ring finger. The MCP, PIP, and DIP joints of the index, middle, and ring fingers appear visually clustered and ambiguous, with minimal differences in their pixel coordinates (Table 3). Consequently, their computed 3D spatial coordinates converge, appearing closely grouped even for joints not identified as outliers. This clustering of ambiguous data is illustrated in Figure 8b.

As shown in Table 4, five landmarks associated with the joints are detected as missing, namely the DIP joint of the middle finger and all four joints of the ring finger. Additionally, the index finger's PIP, DIP, and tip joints are not detected as outliers, due to their near-identical values. This observation can further demonstrate the need for using the kinematic model constraints within the proposed estimation techniques of this study to distinguish individual joints accurately under occlusion conditions. By constraining finger lengths and estimating missing values, a realistic representation of finger poses can be restored despite the occlusion-induced data loss, thereby maintaining consistency with the hand's natural structure.

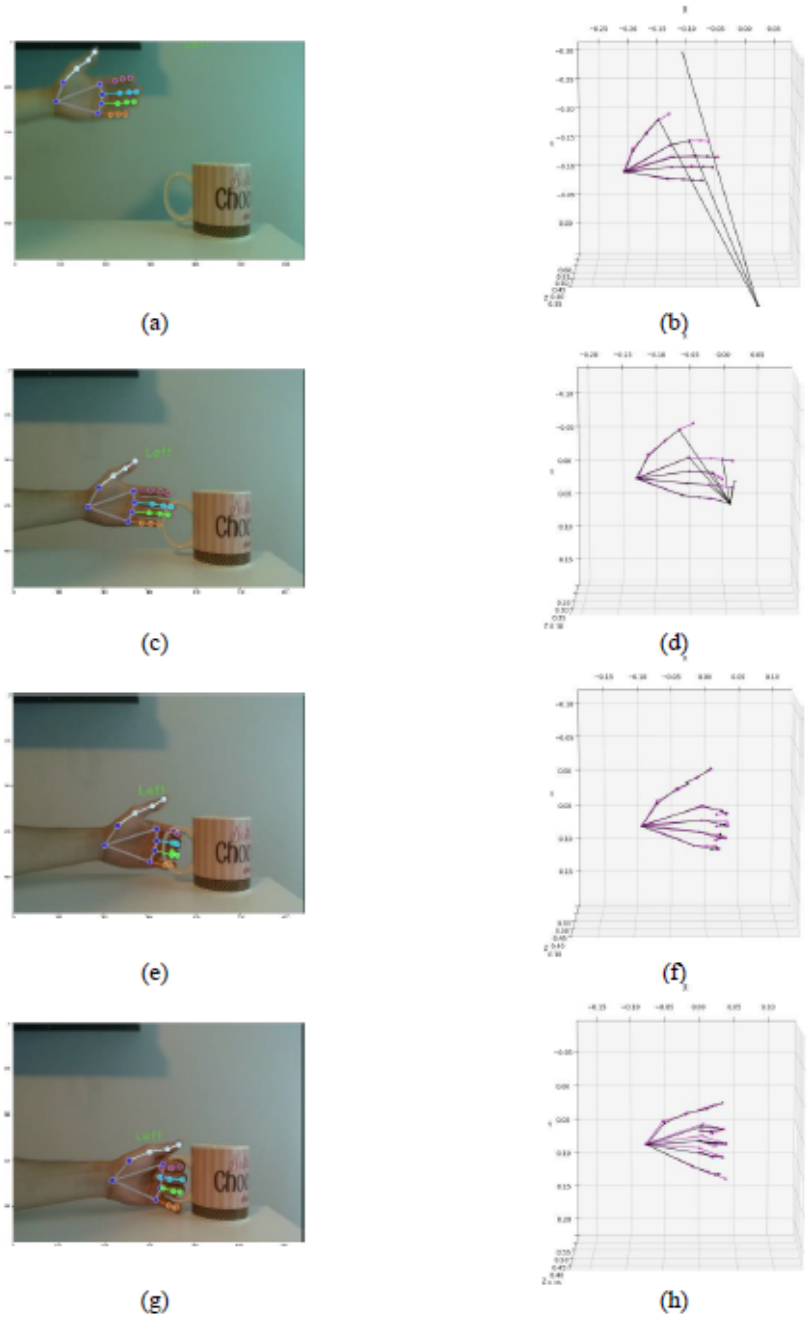

Figure 10. Hand tracking sequences in the frontal view: (a) α Instant – landmarks, (b) α Instant – measurements, (c) β Instant – landmarks, (d) β Instant – measurements, (e) γ Instant – landmarks, (f) γ Instant – measurements, (g) δ Instant – landmarks, (h) δ Instant - measurements

Furthermore, it can be seen that in the absence of the root joint for the ring finger, determining the hand palm as a planar reference becomes infeasible. This lack of a reference plane disrupts the kinematic calculations, leading to failures in most parts of the model due to missing values and instances of self-occlusion. In these scenarios, multiple points converge spatially, resulting in nearly indistinguishable joint angles, which hinders accurate angle calculations. As seen in the unity reconstruction (Figure 8c), these limitations manifest as an unnatural hand orientation, with almost all fingers appearing unnaturally compressed together. This distortion occurs because the lack of critical reference points and reliable spatial separation between joints leaves the kinematic model with insufficient information to accurately resolve individual finger positions and orientations. Consequently, the constraints typically applied to enforce anatomical structure are ineffective, causing fingers to "collapse" into each other.

The inset focuses on iterations 5 to 36, showing gradual loss reduction after the initial sharp decline.

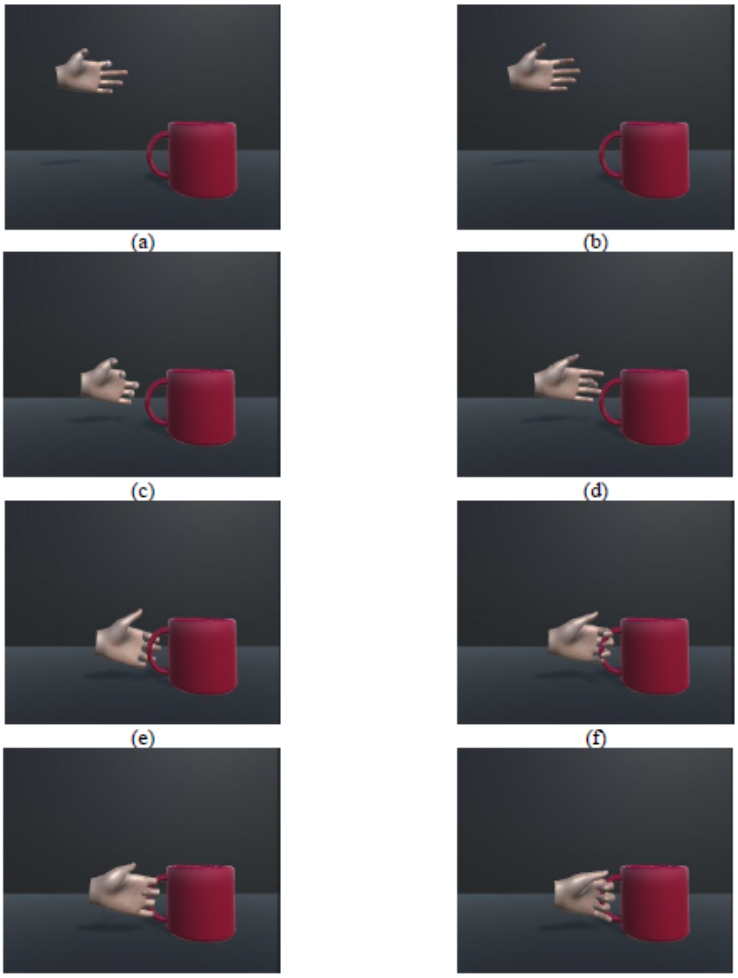

Figure 11. Unity hand graphical reconstruction associated with the sequences of Figure 10: (a) α Instant – original, (b) α Instant – refined, (c) β Instant – original, (d) β Instant – refined, (e) γ Instant – original, (f) γ Instant – refined, (g) δ Instant – original, (h) δ Instant - refined

Table 3. 2D pixel coordinates p_i of 21 MediaPipe hand landmarks in the red–green–blue (RGB) image

	Joints			
Finger	**Root/MCP**	**PIP**	**DIP**	**Tip**
Wrist	$p_0 = (301, 242)$			
Thumb	$p_1 = (276, 250)$	$p_2 = (245, 246)$	$p_3 = (216, 250)$	$p_4 = (196, 254)$
Index	$p_5 = (266, 224)$	$p_6 = (257, 208)$	$p_7 = (250, 206)$	$p_8 = (241, 209)$
Middle	$p_9 = (292, 218)$	$p_{10} = (291, 197)$	$p_{11} = (287, 192)$	$p_{12} = (283, 199)$
Ring	$p_{13} = (317, 219)$	$p_{14} = (323, 200)$	$p_{15} = (328, 200)$	$p_{16} = (329, 205)$
Little	$p_{17} = (337, 225)$	$p_{18} = (356, 210)$	$p_{19} = (371, 210)$	$p_{20} = (382, 217)$

To achieve optimal parameter values of Θ, the numerical optimization methods, highlighted in the previous section, were used to minimize the objective function in Eq. (8), refining the estimated parameters to closely fit the observed hand motions while preserving biomechanical constraints. As illustrated in Figure 7, the iterative optimization loop adjusts the parameters of the kinematic model to reduce estimation errors. Initially, the spatial coordinates ($\mathbf{d}_i^w$ for $i = 0, \ldots, 20$) are obtained from MediaPipe and the depth sensor, with the link lengths $l_{i,j}$. The optimization proceeds by evaluating the objective function (or "loss") and adjusting the model parameters using gradient-based methods. To analyze the performance of the proposed estimation algorithm, the objective function's decreasing trend is plotted in Figure 12. By examining the error reduction trend in Figure 12, substantial decreases in error across key iterations (from iteration 0 to 1,1 to 5, and 5 to the final iteration at 36) can be observed, with reductions of 0.35, 0.07, and 0.002, respectively. These reductions indicate clear convergence toward minimizing error as the algorithm iteratively adjusts parameters, refining the estimated poses to better align with the measured data.

Through iterative refinement, the optimization process successfully restores anatomical coherence to the hand model by adjusting joint orientations and positions, effectively compensating for the missing data. This outcome underscores the robustness of the estimation method in dealing with occlusions and highlights the value of iterative optimization for achieving realistic motion representation. The estimated 3D spatial coordinates of the missing landmarks, along with a comparative analysis of Unity reconstructions with and without the application of the estimation process, are presented in Table 5 and Figures 8c-d.

Table 4. Measured 3D spatial coordinates $\mathbf{d}_i^w$ of 21 hand landmarks (in meters)

	Joints			
Finger	**Root/MCP**	**PIP**	**DIP**	**Tip**
Wrist	(-0.02, -0.01, 0.57)			
Thumb	(-0.04, 0.00, 0.57)	(-0.07, -0.00, 0.56)	(-0.09, 0.00, 0.54)	(-0.11, 0.01, 0.53)
Index	(-0.05, -0.02, 0.52)	(-0.05, -0.03, 0.46)	(-0.05, -0.03, 0.45)	(-0.06, -0.03, 0.44)
Middle	(-0.03, -0.02, 0.51)	(-0.02, -0.04, 0.44)	(0.00, 0.00, 0.00)	(-0.03, -0.03, 0.43)
Ring	(0.00, 0.00, 0.00)	(0.00, 0.00, 0.00)	(0.00, 0.00, 0.00)	(0.00, 0.00, 0.00)
Little	(0.00, 0.00, 0.00)	(0.03, -0.03, 0.50)	(0.04, -0.03, 0.48)	(0.05, -0.02, 0.47)

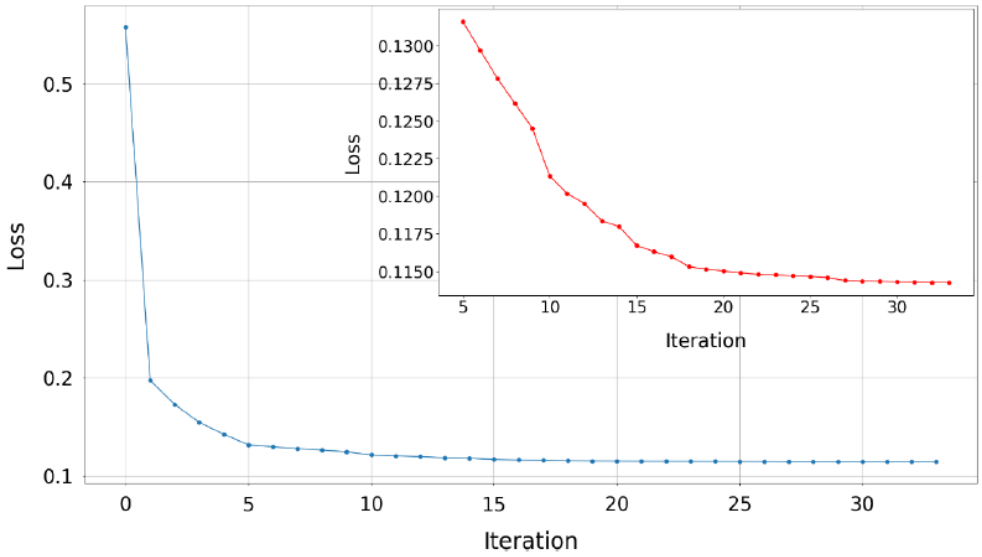

Figure 12. Loss function convergence with a rapid drop in the first iteration

Table 5. Estimated 3D spatial coordinates $\mathbf{d}_i^w$ of 21 hand landmarks (in meters)

	Joints			
Finger	**Root/MCP**	**PIP**	**DIP**	**Tip**
Wrist	(-0.02, -0.00, 0.59)			
Thumb	(-0.04, -0.01, 0.56)	(-0.07, -0.00, 0.55)	(-0.10, 0.00, 0.54)	(-0.12, 0.01, 0.53)
Index	(-0.05, -0.02, 0.51)	(-0.05, -0.03, 0.47)	(-0.05, -0.03, 0.45)	(-0.05, -0.03, 0.44)
Middle	(-0.03, -0.03, 0.51)	(-0.03, -0.04, 0.47)	(-0.03, -0.04, 0.43)	(-0.03, -0.04, 0.43)
Ring	(-0.01, -0.03, 0.51)	(0.00, -0.03, 0.48)	(0.00, -0.04, 0.46)	(0.01, -0.03, 0.44)
Little	(0.03, -0.03, 0.52)	(0.04, -0.03, 0.50)	(0.04, -0.03, 0.48)	(0.05, -0.02, 0.47)

4.2 γ Instant of the Side Measure: Case 2

This instant of the hand-grasping phase is shown in Figure 10. This instant would be the last phase of the grasping hand approaching the cup handle while the tracking information is collected from the side using the RGB-D sensor. In this fully grasping case, all four fingers rotate backward in the direct graphical reconstruction, with each joint-MCP, PIP, DIP, and tip-approaching a 90-degree rotation. This scenario presents a challenging case of object occlusion, as the MCP joints of all four fingers are obscured by the cup handle and the other finger segments. Additionally, the

measured data from the depth sensor becomes nearly collinear, which deviates from the true finger configuration due to the imposed joint angle constraints limiting rotations between 0 and 90 degrees. It can be seen that despite the dominant presence of ambiguities in the landmark measured which resulted in the infeasible joint angles, the proposed method offers a robust joint angle estimation. By incorporating prior knowledge, kinematic constraints, and initial joint angle measures, it can compensate for the collinearity and occlusion-induced errors (Figure 13). The model's optimization process handles the unnatural alignment suggested by the collinear data and adjusts the estimated joint positions accordingly, preserving the realistic articulation of the fingers. This allows us to maintain a coherent representation of the hand pose, even under the presence of object occlusion, as shown in Figure 10d and Figure 14.

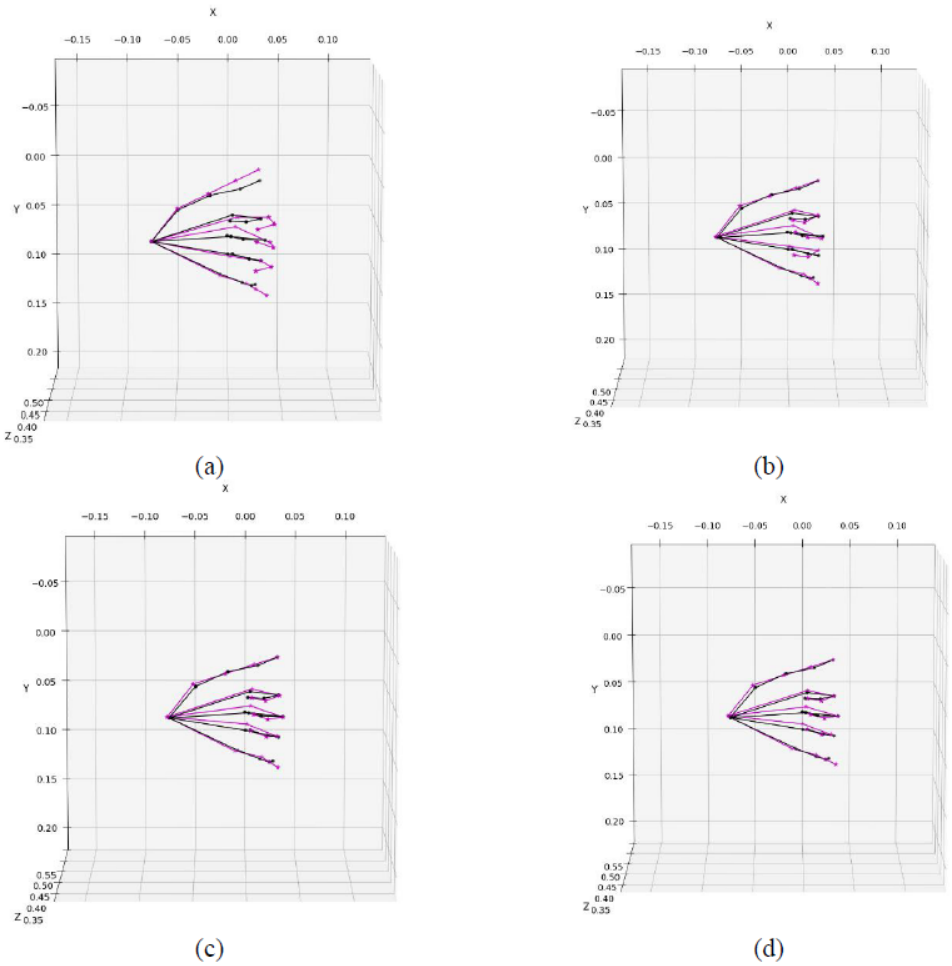

Figure 13. Comparison of estimated and measured hand poses at selected iterations for a fully grasping hand: (a) iteration 0; (b)iIteration 1; (c) iteration 2; (d) final iteration

Figure 14. Comparison of unity hand motion reconstructions during grasping: (a) reconstruction using proposed estimation method; (b) without inclusion of estimation

5 Conclusion

The proposed estimation method demonstrates considerable robustness in handling occlusions, noise, and incomplete data during hand pose tracking and grasping tasks. By leveraging a Bayesian network model, a structured framework was established that captures the hierarchical dependencies among hand joints, allowing the system to effectively compensate for missing depth values and occluded joints. This approach, combined with kinematic constraints, enables accurate hand pose estimation even in complex scenarios (e.g., occlusions caused by object grasping) where sensor data may become unreliable or collinear. The integration of prior knowledge and structural constraints enables the model to produce reasonable approximations despite challenging data conditions. A key limitation of the current method lies in its computational cost. While the Bayesian network is effective at handling noise and missing data, it depends on an accurate understanding of the kinematic structure and inter-joint dependencies to function optimally. The reliance on MLE and iterative optimization techniques introduces computational overhead, which poses challenges for real-time applications that demand rapid processing.

In the current setup, object detection and MediaPipe hand tracking operate efficiently, with execution times of approximately 10 ms and 60 ms per frame, respectively. This efficiency is largely attributed to the extensive training data and optimized model architecture underlying MediaPipe, enabling robust and fast performance. In contrast, the MLE-based parameter estimation is significantly more time-consuming, requiring between 1.5 to 6 seconds per frame. This delay arises from the complexity of aligning point cloud data with the kinematic model, as well as the iterative optimization required to estimate joint positions accurately. Each step involves fine-tuning parameters to satisfy hierarchical constraints while fitting the observed data, making the process computationally intensive. Furthermore, the RGB-D sensor used in the proposed setup has a frame rate of approximately 15 frames per second (FPS), underscoring the mismatch between data acquisition speed and the processing speed of the proposed estimation pipeline.

To address these limitations, several strategies can be explored to enhance the computational efficiency of the estimation framework. These include the adoption of faster optimization techniques, such as diffusion models, which manage uncertainty through probabilistic distributions and have shown promising performance in 3D generative tasks, including unseen point cloud generation and missing part completion. Incorporating prior knowledge through deep learning methods, such as selective optimization guided by learned affordances (e.g., Affordance Diffusion), may accelerate estimation by filtering out inaccurate landmark detections. In addition, hardware acceleration strategies (e.g., edge computing) can enable parallel computations across distributed nodes, further improving efficiency. Together, these advancements present a promising pathway for overcoming current computational bottlenecks and achieving real-time hand pose estimation, thereby increasing the system's practicality and enabling further comparative studies with existing methods.

Author Contributions

Conceptualization, S.P.; methodology, S.P. and Y.Y.D.; software, Y.Y.D.; validation, Y.Y.D.; resources, S.P.; original draft preparation, S.P. and Y.Y.D.; writing — review and editing, S.P. and Y.Y.D.; visualization, Y.Y.D.; supervision, S.P.; funding acquisition, S.P. All authors have read and agreed to the published version of the manuscript.

Funding

This research was funded through the Natural Sciences and Engineering Research Council of Canada.

Data Availability

The data used to support the research findings are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] A. Sadeghi-Niaraki and S. Choi, "A survey of marker-less tracking and registration techniques for health and environmental applications to augmented reality and ubiquitous geospatial information systems," *Sensors*, vol. 20, no. 10, p. 2997, 2020. https://doi.org/10.3390/s20102997

[2] A. Ahmad, C. Migniot, and A. Dipanda, "Tracking hands in interaction with objects: A review," in *2017 13th International Conference on Signal-Image Technology and Internet-Based Systems (SITIS), Jaipur, India*, 2017, pp. 360–369. https://doi.org/10.1109/sitis.2017.66

[3] A. Erol, G. Bebis, M. Nicolescu, D. Richard Boyle, and X. Twombly, "Vision-based hand pose estimation: A review," *Vis. Image Underst.*, vol. 108, no. 1-2, pp. 52–73, 2007. https://doi.org/10.1016/j.cviu.2006.10.012

[4] J. Romero, H. Kjellström, and D. Kragic, "Hands in action: Real-time 3D reconstruction of hands in interaction with objects," in *2010 IEEE International Conference on Robotics and Automation, Anchorage, AK, USA*, 2010, pp. 458–463. https://doi.org/10.1109/robot.2010.5509753

[5] E. Theodoridou, L. Cinque, F. Mignosi, G. Placidi, M. Polsinelli, R. Joao Manuel Tavares, and M. Spezialetti, "Hand tracking and gesture recognition by multiple contactless sensors: A survey," *IEEE Trans. Human-Mach. Syst.*, vol. 53, no. 1, pp. 35–43, 2023. https://doi.org/10.1109/thms.2022.3188840

[6] R. James Carey, L. Tanya Baxter, and P. Richard Di Fabio, "Tracking control in the nonparetic hand of subjects with stroke," *Arch Phys. Med. Rehabil.*, vol. 79, no. 4, pp. 435–441, 1998. https://doi.org/10.1016/s0003-9993(98)90146-0

[7] X. Liang, E. Kapetanios, B. Woll, and A. Angelopoulou, "Real time hand movement trajectory tracking for enhancing dementia screening in ageing deaf signers of British sign language," *IFIP Adv. Inf. Commun. Technol.*, pp. 377–394, 2019. https://doi.org/10.1007/978-3-030-29726-8_24

[8] C. D. Hayden, B. P. Murphy, O. Hardiman, and D. Murray, "Measurement of upper limb function in ALS: A structured review of current methods and future directions," *J. Neurol.*, vol. 269, no. 8, pp. 4089–4101, 2022. https://doi.org/10.1007/s00415-022-11179-8

[9] R. Y. Wang and J. Popović, "Real-time hand-tracking with a color glove," *ACM Trans. Graph.*, vol. 28, no. 3, pp. 1–8, 2009. https://doi.org/10.1145/1531326.1531369

[10] G. Buckingham, "Hand tracking for immersive virtual reality: Opportunities and challenges," *Front. Virtual Real.*, vol. 2, p. 728461, 2021. https://doi.org/10.3389/frvir.2021.728461

[11] A. Ahmad, C. Migniot, and A. Dipanda, "Hand pose estimation and tracking in real and virtual interaction: A review," *Image Vis. Comput.*, vol. 89, pp. 35–49, 2019. https://doi.org/10.1016/j.imavis.2019.06.003

[12] J. Wang and S. Payandeh, "Hand motion and posture recognition in a network of calibrated cameras," *Adv. Multimed.*, vol. 2017, p. 2162078, 2017. https://doi.org/10.1155/2017/2162078

[13] L. Ge, H. Liang, J. Yuan, and D. Thalmann, "Robust 3D hand pose estimation from single depth images using multi-view CNNs," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4422–4436, 2018. https://doi.org/10.1109/tip.2018.2834824

[14] F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C. L. Chang, and M. Grundmann, "Mediapipe hands: On-device real-time hand tracking," *arXiv preprint arXiv:2006.10214*, 2020. https://doi.org/10.48550/ARXIV.2006.10214

[15] I. Oikonomidis, N. Kyriazis, and A. A. Argyros, "Efficient model-based 3D tracking of hand articulations using Kinect," in *Proc. Brit. Mach. Vis. Conf.*, vol. 2, 2021, pp. 1–11. https://doi.org/10.5244/c.25.101

[16] A. Tagliasacchi, M. Schröder, A. Tkach, S. Bouaziz, M. Botsch, and M. Pauly, "Robust articulated-ICP for realtime hand tracking," *Comput. Graph. Forum*, vol. 34, no. 5, pp. 101–114, 2015. https://doi.org/10.1111/cgf.12700

[17] C. Qian, X. Sun, Y. Wei, X. Tang, and J. Sun, "Realtime and robust hand tracking from depth," in *IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA*, 2014, pp. 1106–1113. https://doi.org/10.1109/cvpr.2014.145

[18] C. Diaz and S. Payandeh, "Multimodal sensing interface for haptic interaction," *J. Sens.*, vol. 2017, no. 1, p. 2072951, 2017. https://doi.org/10.1155/2017/2072951

[19] W. F. B. W. Tarmizi, I. Elamvazuthi, and M. Begam, "Kinematic and dynamic modeling of a multi-fingered robot hand," *Int. J. Basic Appl. Sci.*, vol. 9, no. 10, pp. 89–96, 2009.

[20] D. H. Lee, J. H. Park, S. W. Park, M. H. Baeg, and J. H. Bae, "Kitech-hand: A highly dexterous and modularized robotic hand," *IEEE/ASME Trans. Mechatron.*, vol. 22, no. 2, pp. 876–887, 2017. https://doi.org/10.1109/tmech.2016.2634602

[21] Z. Wang, K. Zhang, and R. Sankaranarayana, "Dm-HAP: Diffusion model for accurate hand pose prediction," *Neurocomputing*, vol. 611, p. 128681, 2025. https://doi.org/10.1016/j.neucom.2024.128681

[22] W. Cheng, H. Tang, L. Van Gool, and J. H. Ko, "Handdiff: 3D hand pose estimation with diffusion on image-point cloud," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA*, 2024, pp. 2274–2284. https://doi.org/10.1109/cvpr52733.2024.00221

[23] Y. Ye, X. Li, A. Gupta, S. De Mellon, S. Birchfield, J. Song, S. Tulsiani, and S. Liu, "Affordance diffusion: Synthesizing hand-object interactions," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada*, 2023, pp. 22 479–22 489. https://doi.org/10.1109/cvpr52729.2023.02153

[24] Y. Ye, P. Hebbar, A. Gupta, and S. Tulsiani, "Diffusion-guided reconstruction of everyday hand-object interaction clips," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France*, 2023, pp. 19 717–19 728.

[25] M. Mangalam, S. Oruganti, G. Buckingham, and W. Christoph Borst, "Enhancing hand-object interactions in

virtual reality for precision manual tasks," *Virtual Real.*, vol. 28, no. 4, 2024. https://doi.org/10.1007/s10055-024-01055-3

[26] H. R. Joseph Isaac, M. Manivannan, and B. Ravindran, "Single shot corrective CNN for anatomically correct 3D hand pose estimation," *Front. Artif. Intell.*, vol. 5, p. 759255, 2022. https://doi.org/10.3389/frai.2022.759255

[27] G. Pavlakos, D. Shan, I. Radosavovic, A. Kanazawa, D. Fouhey, and J. Malik, "Reconstructing hands in 3D with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA*, 2023, pp. 9826–9836. https://doi.org/10.48550/ARXIV.2312.05251

[28] H. Dong, A. Chharia, W. Gou, F. V. Carrasco, and F. De la Torre, "Hamba: Single-view 3D hand reconstruction with graph-guided bi-scanning mamba," *Adv. Neural Inf. Process. Syst.*, vol. 37, pp. 2127–2160, 2024. https://doi.org/10.48550/ARXIV.2407.09646

[29] M. Rezaei, R. Rastgoo, and V. Athitsos, "Trihorn-Net: A model for accurate depth-based 3D hand pose estimation," *Expert Syst. Appl.*, vol. 223, p. 119922, 2023. https://doi.org/10.1016/j.eswa.2023.119922

[30] A. Saran, D. Teney, and M. Kris Kitani, "Hand parsing for fine-grained recognition of human grasps in monocular images," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Hamburg, Germany*, 2015, pp. 5052–5058. https://doi.org/10.1109/iros.2015.7354088

[31] D. Geiger and D. Heckerman, "Learning gaussian networks," in *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, 1994, pp. 235–243.

[32] M. S. Ahn, H. Chae, D. Noh, H. Nam, and D. Hong, "Analysis and noise modeling of the intel realsense D435 for mobile robots," in *Proceedings of the International Conference on Ubiquitous Robots (UR), Jeju, Korea (South)*, 2019, pp. 707–711. https://doi.org/10.1109/urai.2019.8768489

[33] E. Curto and H. Araujo, "Fitting a normal probability distribution to depth estimations of three REALSENSE™ RGB-D cameras tested in scenes with transparency," *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. 48, pp. 59–63, 2022. https://doi.org/https://doi.org/10.5194/isprs-archives-XLVIII-2-W1-2022-59-2022

[34] A. Hald, "On the history of maximum likelihood in relation to inverse probability and least squares," *Stat. Sci.*, vol. 14, no. 2, pp. 214–222, 1999.